DISS. ETH NO. 19878

DATA MINING AND DATA-DRIVEN MODELING APPROACHES TO SUPPORT WASTEWATER TREATMENT PLANT OPERATION

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

DAVID JÉRÔME DÜRRENMATT

Dipl. Umwelt.-Ing. ETH

born October 30, 1981

citizen of Guggisberg (BE) and Bassersdorf (ZH)

accepted on the recommendation of

Prof. em. Dr. Willi Gujer, examiner

Prof. Dr. Eberhard Morgenroth, co-examiner

Assoc. Prof. Dr. Gürkan Sin, co-examiner

2011

Essentially, all models are wrong, but some are useful.

— *George E. P. Box*

Abstract

In wastewater treatment plants (WWTPs), much effort and money is invested in operating and maintaining dense plant-wide measuring networks. The network primarily serves as input for the advanced control scenarios that are implemented in the supervisory control and data acquisition (SCADA) system to satisfy the stringent effluent quality constraints. Due to new developments in information technology, long-term archiving has become practicable, and specialized process information systems are now available. The steadily growing amount of plant data available, however, is not systematically exploited for plant optimization because of the lack of specialized tools that allow operators and engineers alike to extract meaningful and valuable information efficiently from the massive amount of high-dimensional data. As a result, most information contained in the data is eventually lost.

In the past few years, many data mining techniques have emerged that are capable of analyzing massive amounts of data. Available processing power allowed the development of efficient data-driven modeling techniques especially suited to situations in which the speed of data acquisition surpasses the time available for data analysis. However, although these methods are promising ways to provide valuable information to the operator and engineer, there is currently no fully developed interest in the application of these techniques to support WWTP operation.

In this thesis, the applicability of data mining and data-driven modeling techniques in the context of WWTP operation is investigated. This context, however, implies specific characteristics that the adapted and developed techniques must satisfy to be practicable: On the one hand, the deployment of a given technique on a plant must be fast, simple and cost-effective. As a consequence, it must consider data that are already available or that can be gathered easily. On the other hand, the application must be safe, i.e., the extracted information must be reliable and communicated clearly. This thesis presents the results of four knowledge discovery projects that adapted data mining and data-driven modeling techniques to tackle problems relevant to either the operator or the process engineer.

First, the extent to which data-driven modeling techniques are suitable for the automatic generation of software sensors exclusively based on measured data available in the SCADA system of the plant is investigated. These software sensors are meant to be substitutes for failure-prone and maintenance-intensive sensors and to diagnose hardware sensors. In two full-scale experiments, four modeling techniques for software-sensor development are compared and the role of expert knowledge is investigated. The investigations show that the non-linear modeling techniques outperform the linear technique and that a higher degree of expert knowledge is beneficial for long term accuracy, but can lead to reduced performance in the short term. Consequently, if frequent model recalibration is possible, as is the case for sensor diagnosis applications, automatic development given limited expert knowledge is feasible. In contrast, optimum use of expert knowledge requires model transparency, which is only given for two of the investigated techniques: generalized least squares regression and self-organizing maps.

In the second project, WWTP operators are provided with additional information on characteristic sewage compositions arriving at their plant from clustered UV/Vis spectra measured at the influent. A two-staged clustering approach is considered that copes well with high-dimensional and noisy data. If it is possible to assign a characteristic cluster to a sewage producer in the catchment, detailed analysis of the temporal discharging pattern is possible without the need for additional measurements at the production site. In a full-scale experiment, one of five detected clusters could by assigned to an industrial laundry by analyzing the cluster centroids. In a validation experiment, 93 out of 95 discharging events were classified correctly. Successful detection depends on the uniqueness of the producer's UV/Vis pattern, the dilution at the influent and the size and complexity of the catchment.

In WWTPs, asymmetric feeding of reactors operating in parallel lanes can lead to operational issues and significant performance losses. A new method based on dynamic time warping is presented that makes the quantification of the discharge distribution at hydraulic flow dividers practicable. The method estimates the discharge distribution as a function of total discharge at the divider given influent and effluent measurements of some measured signal in the downstream reactors. The function can not only serve as the basis for structural modification, but it can also be used to calculate the flow to the individual lanes given the total influent, and thus avoid the assumption of equal distribution (this assumption must often be made by process engineers and scientists). Theoretical analysis reveals that the accuracy of the function depends on the hydraulic residence time, the dispersion and the reactions in the reactors downstream of the divider, in addition to the variability of the signal. A systematic application on a wide range of synthetic systems that may be found on WWTPs shows that the error is at least half that when an equal distribution is assumed if the function is used to obtain a better estimate for the flow to a reactor. In a full scale validation experiment, the discharge distribution could be accurately estimated.

The fourth application presented shows that optimal hydraulic reactor models can be searched automatically using grammar-based genetic programming. This method is especially relevant for engineers who want to model the hydraulic processes of the plant and, because of the limited applicability of existing approaches, must rely solely on their experience and intuition for further insights into the reactor hydraulics. With a tree encoding that can decode program trees into hydraulic reactor models compatible with common software and with influent and effluent measurements, a palette of equally performing models can be generated. Of these the modeler then picks the most suitable one as starting point. The methodology is applied to reverse-engineer synthetic systems, and because of theoretical and practical identifiability issues, several searches yield different models, which emphasizes the need for an expert to choose the most appropriate model. The method is applied to generate reactor models of a primary clarifier with unknown exact volume. The volume of the resulting models corresponds to the expectation and virtual tracer experiment performed on the synthetic models generally confirms with an experiment performed on-site.

The knowledge discovery projects show that optimal model choice and complexity greatly depend on the specific problem and on the degree of available expert knowledge. In general, safe deployment on-site requires transparent models that can be interpreted even with limited knowledge and intuitive and understandable communication of the model results.

Because the effluent quality constraints will further tighten and progress in the fields of information technology and data analysis will continue, it is necessary to use the available data to fully exploit the plants. Data mining and data-driven modeling are suitable tools.

Zusammenfassung

In Kläranlagen wird viel Aufwand und Geld in den Betrieb und die Wartung eines anlagenweiten Messnetzes gesteckt. Dieses Messnetz dient in erster Linie als Input für die Regelungsvorgänge, die im Prozessleitsystem definiert sind, um die strengen Grenzwerte bezüglich der Ablaufqualität zu erfüllen. Neue Errungenschaften im Bereich der Informationstechnologie ermöglichen nun eine wirtschaftliche Langzeitarchivierung und es stehen mittlerweile sogar spezialisierte Prozessinformationssysteme zur Verfügung. Allerdings wird die stetig wachsende Menge an verfügbaren Anlagedaten nicht systematisch ausgenutzt und zur Optimierung der Anlagen herangezogen. Dies ist auf das Fehlen von spezialisierten Instrumenten zurückzuführen, die es sowohl Betreibern als auch Ingenieuren ermöglichten, aussagekräftige und wertvolle Informationen aus der gewaltigen Menge hochdimensionaler Daten zu extrahieren. Folglich geht ein Grossteil der in den Daten enthaltenen Informationen verloren.

In den letzten Jahren wurden zahlreiche Data-Mining-Verfahren entwickelt, die zur Analyse massiver Datenmengen geeignet sind. Die verfügbare Rechenleistung ermöglicht die Entwicklung effizienter, datenbasierter Modellierungstechniken, die sich besonders dann eignen, wenn sich Daten schneller anhäufen, als dass sie analysiert werden können. Obwohl diese Methoden geeignet wären, um Betreiber und Ingenieure mit wertvollen Informationen zur Unterstützung des Kläranlagenbetriebs zu versorgen, ist ihre Anwendung noch nicht etabliert.

Die vorliegende Arbeit untersucht die Verwendbarkeit von Data-Mining und datenbasierter Modellierung zur Unterstützung des Kläranlagenbetriebs. Dieser Anwendungsbereich setzt allerdings besondere Eigenschaften an die entwickelten Methoden voraus, damit sie praktikabel sind. Dies bedeutet zum einen, dass der Einsatz auf einer Anlage schnell, einfach und kostengünstig durchführbar ist. Daraus folgt, dass mit Vorteil Daten berücksichtigt werden, die entweder bereits zur Verfügung stehen oder die leicht gesammelt werden können. Andererseits muss die Anwendung sicher sein, d. h. die gewonnenen Erkenntnisse sollten verlässlich sein und verständlich kommuniziert werden. In dieser Arbeit werden die Ergebnisse aus vier Forschungsprojekten präsentiert, bei denen angepasste Data-Mining und datenbasierte Modellierungstechniken zur Anwendung kommen, um für Betreiber und Ingenieure relevante Probleme zu lösen.

Zunächst wird untersucht, in welchem Ausmass sich die datenbasierte Modellierung für die automatische Generierung von Software-Sensoren, die ausschliesslich auf den im Prozessleitsystem verfügbaren Daten basieren, eignet. Diese Software-Sensoren sollen fehleranfällige und wartungsintensive Sensoren ersetzen und ausserdem zur Diagnose der Hardware-Sensoren verwendet werden können. In zwei grosstechnischen Experimenten wurden vier Modellierungsmethoden für die Entwicklung von Software-Sensoren miteinander verglichen und die Bedeutung von Expertenwissen untersucht. Die Untersuchung zeigt, dass bezüglich Genauigkeit die nicht-linearen Methoden die linearen übertreffen, sowie dass hochgradiges Expertenwissen langfristig eine grössere Genauigkeit gewährleistet, während es die Genauigkeit kurzfristig gesehen reduziert. Ist also eine laufende Rekalibrierung möglich, wie z.B. bei Anwendungen zur Sensordiagnose, ist die automatische Generierung auch bei begrenztem Expertenwissen realisierbar. Im Gegenzug erfordert die optimale Ausnutzung von Expertenwissen Modelltransparenz. Diese ist nur für zwei der untersuchten Methoden gegeben, nämlich für verallgemeinerte Kleinste-Quadrate-Modelle und selbstorganisierende Karten.

Im zweiten Projekt wird aufgezeigt, wie Kläranlagenbetreiber durch Clustering von im Zulauf gemessenen UV/Vis-Spektren zusätzliche Informationen zu charakteristischen Abwasserzusammensetzungen gewinnen können. Dabei wird eine zweistufige Clustering-Methode eingesetzt, die sich besonders für hochdimensionale und verrauschte Daten eignet. Sofern ein Abwasserproduzent im Einzugsgebiet einem Cluster zugeordnet werden kann, ist eine detaillierte Analyse der Einleitvorgänge auch ohne zusätzliche Messung beim Produzenten möglich. Im Rahmen eines grosstechnischen Experiments ist es gelungen, einen von fünf entdeckten Clustern durch Analyse der Clusterschwerpunkte und Einleitungsmuster einer Grosswäscherei zuzuordnen. Bei einem Validierungsexperiment wurden 93 von 95 Einleitungen richtig zugeordnet. Die erfolgreiche Zuordnung hängt von der Besonderheit des UV/Vis-Spektrums des Produzenten, seiner Verdünnung im Zulauf und der Komplexität des Einzugsgebiets ab.

In Kläranlagen kann die ungleichmässige Beschickung mehrerer parallel betriebener Strassen zu Betriebsproblemen und Leistungseinbussen führen. Im dritten Projekt wird deshalb eine neue Methode vorgestellt, die auf Dynamic Time Warping basiert, und die die Quantifizierung der Durchflussverteilung in hydraulischen Trennbauwerken praktikabel macht. Diese Methode schätzt eine Funktion, die die Verteilung in Abhängigkeit zum gesamten Durchfluss beschreibt. Sie setzt lediglich die Messung eines fast beliebigen Signals in den Zu- und Abläufen der nachgeschalteten Reaktoren voraus. Die geschätzte Funktion kann einerseits als Grundlage für bauliche Anpassungen dienen. Andererseits kann sie aber auch verwendet werden, um den Durchfluss in die einzelnen Reaktoren bei bekanntem Gesamtdurchfluss zu bestimmen und so die Annahme gleichmässiger Beschickung zu vermeiden. Eine theoretische Analyse zeigt, dass die Genauigkeit von der hydraulischen Verweilzeit, der Dispersion und den Reaktionen in den Reaktoren sowie von der Variabilität des gemessenen Signals abhängt. Die Anwendung in verschiedenen synthetischen Systemen zeigt, dass der Fehler im Vergleich zur Annahme gleichmässiger Beschickung mindestens halbiert werden kann, wenn der Durchfluss mit der geschätzten Funktion bestimmt wird. Die Durchflussverteilung konnte in einem Validierungsexperiment mithilfe der beschriebenen Methode genau bestimmt werden.

Die vierte Anwendung schliesslich zeigt, dass mithilfe grammatikbasierter genetischer Programmierung automatisch hydraulische Reaktormodelle realer Reaktoren gefunden werden können. Dies ist besonders für Ingenieure relevant, die die hydraulischen Prozesse einer Kläranlage modellieren möchten und sich oft auf ihre Erfahrung und Intuition verlassen müssen, da existierende Verfahren zur Bestimmung der Reaktorhydraulik zu aufwändig wären. Mit einer Kodierung, die Programme in Reaktormodelle übersetzt sowie Messungen im Zu- und Ablauf des zu modellierenden Reaktors kann in mehreren Läufen eine Auswahl an passenden Modellen erzeugt werden. Aus dieser wählt der Modellierer anschliessend das am besten geeignete Modell aus. Die Anwendung der Methode zum Nachbau künstlicher Systeme zeigt, dass in mehreren Läufen erzeugte Modelle aufgrund theoretischer und praktischer Identifizierbarkeitsgrenzen unterschiedlich sein können. Deshalb ist Expertenwissen zur Wahl des passendsten Modells unverzichtbar. In einem Experiment ist die Methode zur Modellierung eines Vorklärbeckens mit nicht genau bekanntem Volumen erfolgreich angewendet worden.

Die Forschungsprojekte verdeutlichen, dass die optimale Modellierungstechnik und Modellkomplexität von der jeweiligen Anwendung und dem verfügbaren Expertenwissen abhängen. Allgemein erfordert ein sicherer Einsatz am Standort transparente Modelle, die auch mit wenig Wissen interpretiert werden können und die Ergebnisse verständliche kommunizieren.

Weil die Auflagen für die Abwasserqualität in Zukunft weiter verschärft werden und in den Bereichen Informationstechnologie und Datenanalyse mit weiteren Fortschritten gerechnet werden kann, ist es lohnenswert, die verfügbaren Daten in Kombination mit Data-Mining und datenbasierte Modellierung zur Anlagenoptimierung zu nutzen.

Acknowledgements

This thesis would not have been possible without the contribution, advice, comments and motivation I received from many people.

First, I want to express sincere thanks to my supervisor Willi Gujer. He gave me freedom to choose the research topics I wanted to focus on in this project. Whenever I needed somebody to critically comment on my ideas, to encourage and inspire me and to keep me heading in the right direction whenever I was lost in the massive amount of data I tried to mine, he immediately took the time to have a discussion with me, despite his other priorities.

I want to express my gratitude to my co-examiners, Eberhard Morgenroth and Gürkan Sin, for all of their constructive inputs and stimulating discussions.

The staff of WWTP Kloten/Opfikon is gratefully acknowledged for allowing me to perform most of my experiments at their site, to use all of their facilities, to allow me open access to their data and, last but not least, for providing me with the WWTP operator's perspective on my research.

I am very grateful to my friends and colleagues from both the ENG and SWW departments at Eawag. Whether for their academic inputs or simply for a fair amount of aimless talking, they provided the fruitful atmosphere needed to let my ideas grow. The regular UWE meetings provided excellent opportunities to brainstorm and discuss new ideas and to find solutions to current problems by combining the different perspectives of the group members.

My past and current office mates from BU B05 are magnificent fellows. I very much enjoyed their company and I apologize for all my (annoying) moods and the continuous noise of the brewing coffee machine whenever I was in the office. Thank you Sarina Jenni, Bettina Sterkele, Christoph Egger, Markus Gresch, Marc Neumann and Pascal Wunderlin.

Finally, I would like to thank my family for their continuous support during my PhD. Eva, very special thanks are due to you: You quietly accepted the many late hours and weekends in front of the computer and still cheered me up in difficult times. Luckily, it seems to me, you eventually developed a tremendous (for a historian) interest for the field of wastewater treatment and enjoyed our sunny Sunday cycle rides to the WWTP Kloten/Opfikon.

Contents

Chapter 1

Introduction

Introduction

Increasing effluent quality requirements and fiscal restraints force wastewater treatment plant (WWTP) operators to fully exploit the available facilities. For this reason, advanced process control strategies are implemented that require the availability of a dense plant-wide measuring network that consists of a multitude of online and offline sensors. Sensor measurements are primarily used for control. Due to new developments in the field of information technology, however, it is now possible to archive the measured data for an indefinite time in the supervisory control and data acquisition system (SCADA system) of the plant, or in a dedicated process information system.

Although there is a steadily growing amount of historical data at the operator's and engineer's disposal, most information contained in the data will remain unused. From the operator's point of view, the primary reason for this lies in the high dimensionality of the data with which available traditional statistical tools and visualization techniques cannot cope. Consequently, the information contained in the data is lost. The availability of methods and tools that enable systematic extraction of information hidden in the data, however, would assist the operator in further optimization of his or her plant, eventually helping, e.g., to further increase the effluent quality, to reduce the consumption of energy and other resources and to foster the operator's knowledge on the plant processes.

From the engineer's point of view, available historic data collected during routine operation is not sufficient for plant modeling (Dold *et al.*, 2010). Because supplementary sampling is costly and time consuming, additional measuring campaigns are often not realized (Dold *et al.*, 2010). It is often claimed that data collection is one of the main obstacles in the modeling procedure (Hauduc *et al.*, 2009). The application of modern methods and tools for data analysis, however, could not only help to maximize the use of the available data but also help design inexpensive and simple measuring campaigns.

In the last several decades, as a response to the general trend that the amount and complexity of available data are growing faster than the ability to analyze it, data mining and data-driven modeling techniques have been developed. Both techniques are based on data and are aimed at information extraction and, ultimately, knowledge generation; they rely on the availability of computational power. Within an appropriate framework, model generation and deployment can widely be automated and thus provide a cost-effective alternative to prevalent approaches, motivating the further use of the data. In the field of urban water engineering, there is currently no fully developed interest in data mining and data-driven modeling to support WWTP operation. Hence, there exists a niche for automated and modular modeling techniques that are based on either easy-to-measure data or routine data already available that can be deployed easily and reliably and to support WWTP operation.

In this introductory chapter, a short description of data mining and data-driven modeling is first given. Then, common data mining techniques are described, a process model that serves as a guideline for knowledge discovery projects is presented and major challenges are

addressed. The state of the application of data mining and data-driven modeling in the context of wastewater treatment operation is discussed, followed by the definition of the goals and general research questions. An outline of this thesis is included as well.

1.1 Data mining and data-driven modeling

Briefly, data mining is the application of specific algorithms to extract patterns from data (Fayyad *et al.*, 1996).

In the last several decades, data accumulation has become easier, and storage has become inexpensive, while the human processing level has stayed almost constant (Maimon and Rokach, 2005). Traditional methods applied to turn data into knowledge heavily rely on manual analysis and interpretation and are therefore impractical for massive data sets (Fayyad *et al.*, 1996).

Data mining is the response to this technological gap and has an interdisciplinary nature: it encompasses a variety of methods from the overlapping fields of, e.g., statistics, machine learning, pattern recognition and artificial intelligence. Although there is no precise definition of the term "data mining", there is some sort of common understanding of its purpose, which is the use of (novel) methods to analyze large amounts of data (Fayyad *et al.*, 1996). Two additional characteristics are distinguishing for data mining. First, instead of collecting data for the purpose of a particular experiment, the focus is on data that is already available in, e.g., spreadsheets, databases and process information systems. Second, data mining is data-driven. Thus, instead of finding the smallest data set that yields sufficiently confident estimates given a model, it is intended to find a good model that is still easy to understand given a large amount of data (Cios *et al.*, 2007b).

Data mining is viewed as the key phase in the broader KDD process (knowledge discovery in databases). The phrase KDD explicitly emphasizes that the end product of data-driven discovery is knowledge (Fayyad *et al.*, 1996). By defining the main tasks of a knowledge discovery project, KDD ensures that meaningful knowledge is derived. A process model that implements the particular steps of the KDD will be given in Section 1.1.2.

In data-driven modeling, data characterizing a system are analyzed to look for connections between the system state variables without taking into account explicit knowledge of the physical behavior of the system. This approach is in contrast to physically based (or knowledge-driven) modeling, where the aim is to describe the mechanistic behavior of the system (Solomatine and Ostfeld, 2008; Solomatine *et al.*, 2008). Data-driven modeling intersects with the interdisciplinary areas of the fields of data mining, machine learning, statistics, etc. and has a similar focus. However, in contrast to data mining, data-driven modeling does not focus on large data bases and the analysis of secondary data, i.e., data that are already available and thus have not been acquired for a particular experiment. Rather, data-driven modeling is best implemented when it can be based on the use of inexpensive, basic measurement signals to produce parsimonious models that have good generalization ability (Chan, 2003; Dewasme *et al.*, 2009). Obviously, the given discrimination is fuzzy, and often, the same methods can be applied for either data mining or data-driven modeling.

1.1.1 Overview of the data mining techniques

The classification of the data mining techniques given in this section closely follows the suggestions of Fayyad *et al.* (1996). Although there are several suggestions on how to classify the available methods, they generally agree with that publication.

First, it is distinguished whether the primary goal is prediction or description. "Prediction problems" encompass the group of problems whose goal is the development of a mathematical model that is used to predict the expected output, given the input. The performance of such a model can be assessed by evaluating its predictive power. "Description problems", on the other hand, try to find human interpretable patterns that describe the data. The quality of a discovered pattern is assessed by analyzing the descriptive accuracy of a pattern. The boundaries between prediction and description problems, however, are not sharp.

Classification is a predictive method. A learned function is applied to a data item to assign it one of several predefined classes. A typical classification problem could read as follows: given a set of (influent) measurements and the information about whether the effluent concentration constraints are violated or not, is it possible to design a function (given the available data) that can predict the right class for a new measurement? To assess the performance of the function, the classification error can be assessed (i.e., the ratio of false positives (type I error) and false negatives (type II error)).

Regression is a popular predictive method that maps a data item to a real-valued prediction variable, given a learned function. Software sensors, i.e., a piece of software that outputs a quantitative signal based on a model and one or more input signals, for instance, perform a regression task. The quality of the fit is often assessed as a function of the deviations of the model output from measurements.

Clustering, a descriptive task, tries to identify a set of clusters that describe the data. Clustering is an unsupervised method, i.e., no class membership is assumed to be known. As an example, given a set of vectors $v_t = (m_1, m_n, m_N)^T$ that contain N measurements of sensors $n = 1 \ldots N$ at different times t, are there clusters that are typical for certain process states or environmental conditions?

Summarization aims at finding a compact description for the data. Simple examples of summarization include the computation of the mean and standard deviation of the variables.

In *dependency modeling*, one attempts to find a model that describes the significant relationships between variables. Given a set of discharge and rainfall measurements, for instance, dependency modeling could reveal that rainfall coincidences with high discharge values with a certain confidence.

Change and deviation detection tries to discover significant changes in data when compared with previous values. A simple example is the detection of a shift of the mean of a data series.

Obviously, some predictive models can also be descriptive, and vice versa. In Chapter 3, for instance, a two-staged clustering algorithm is first applied to cluster noisy UV/Vis spectra. Then, the resulting clusters are labeled and afterwards used to classify new spectra.

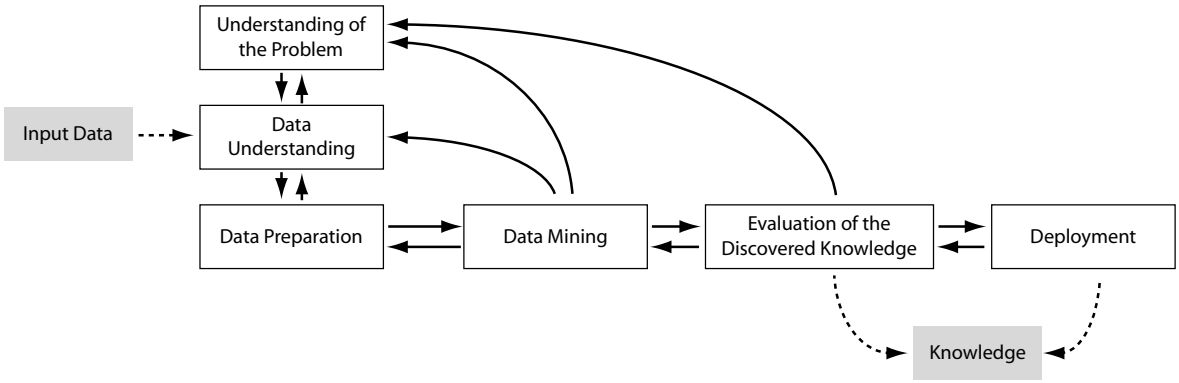

Figure 1.1: Knowledge discovery process model based on the model of Cios *et al.* (2007a).

1.1.2 Workflow of a knowledge discovery process

For a strategic, target-oriented proceeding and to ensure a high chance of the best outcome, it is important to understand the overall approach. For this reason, frameworks that formalize the knowledge discovery process have been introduced (Cios *et al.*, 2007a). These process models define the life cycle of the knowledge discovery project and provide a roadmap that can be followed to execute a project in an arbitrary domain.

The process model applied to the projects presented in this thesis is depicted in Figure 1.1. It is based on the model defined by Cios *et al.* (2007a), which is a hybrid between the industry-oriented CRISP-DM (CRoss-Industry Standard Process for Data Mining) introduced by a consortium of four large European companies (Chapman *et al.*, 2000) and the academic research model provided by Fayyad *et al.* (1996). It consists of the following six connected highly interactive and iterative phases:

i) *Understanding of the problem.* This step includes the definition of the problem and the determination of the objectives. The tools that will be used for data mining are chosen. Important factors that can influence the outcome will be uncovered here. Neglecting this step may result in spending a great deal of effort obtaining the right answers to the wrong question.

ii) *Data understanding.* Here, the data are collected, checked and integrated, possibly by taking into account background knowledge. The usefulness of the data with regard to the objectives is verified.

iii) *Data preparation.* In this phase, which data will be used and in what form is determined. Consequently, significance testing, data cleaning, deriving new attributes and feature selection and extraction are part of this phase. The data are now available in a form that is compatible with the tools selected in the first step.

iv) *Data mining.* Various methods are applied to extract knowledge from the preprocessed data. Extracted knowledge can be of arbitrary form, e.g., a set of rules or a model. Accuracy and generality are assessed in this phase.

v) *Evaluation of the discovered knowledge.* The results are interpreted. It is noted whether there are novel and interesting patterns. Taking into consideration the domain knowl-

edge, the impact of the new knowledge is evaluated. It is determined whether the initially stated goals have been met.

vi) *Deployment.* This step deals with the determination of a deployment strategy: Where and how should the acquired knowledge be used?

Considering the relative effort required for each of these phases, several estimates have been proposed by both researchers and practitioners (Cios *et al.*, 2007a). Roughly 50% of the time is spent on data preparation, which is thus the costliest phase. The other phases demand approximately the same amount of effort.

1.1.3 Application issues

Selecting potential applications depends on practical and technical criteria (Fayyad *et al.*, 1996). Practical criteria include the possible impact of the application, the absence of simpler alternative solutions and organizational support for using a technology. Technical criteria include the availability of sufficient data with relevant attributes (i.e., they must be relevant to the discovery task). An indispensable criterion, however, is the availability of sufficient domain knowledge.

During a knowledge discovery project, the data miner must cope with various application challenges. In Fayyad *et al.* (1996), in a nonexhaustive list, the authors enumerate typical problems that practitioners must deal with. Those that apply to a high degree to the systems considered in this thesis are:

i) *Databases are becoming larger,* not only in terms of the number of records but also in terms of attributes[1]. Hence, there is a need for more efficient algorithms, sampling, approximation and parallel processing.

ii) A *high-dimensional data set*, i.e., a data set with many attributes, on the one hand, increases the search space for model induction and, on the other hand, leads to a phenomenon referred to as the "curse of dimensionality" (Verleysen and François, 2005): The higher the dimensionality, the more equidistant the data points are. Mitigation strategies include the reduction of the dimensionality and the inclusion of prior knowledge to remove irrelevant variables.

iii) *Non-stationary data* can make previously discovered patterns (or models) invalid. This is particularly an issue for the considered systems because the processes and the environmental conditions are constantly subject to change. Possible solutions include frequent updating of the patterns or the consideration of adaptive models.

iv) *Missing and noisy data* can be cleaned with filtering and outlier detection strategies.

v) *Overfitting*, thus the lack of generalization ability of a model, results in poor performance if the model is applied to new data. Cross-validation and other (model-dependent) strategies can be applied to prevent overfitting.

[1] In a database table, a record (also called row) represents a single data item that has a set of attributes (stored in the data table columns).

vi) *Discovered patterns must be made understandable by humans* and effectively communicated. A plethora of data and visualization methods are available for this purpose.

vii) The *inclusion of prior knowledge* in a simple way is not possible for many current methods and tools, although consideration of this knowledge is important for the success of the project.

viii) *Integration with other systems*, such as existing process control systems, is crucial; a stand-alone discovery system is not desired in most cases.

With regard to the knowledge discovery process, list items *(i)–(iv)* are connected to input data, item *(v)* is connected primarily to data mining and items *(vi)–(viii)* are linked to evaluation and deployment.

Basically, the same challenges apply to data-driven modeling except, for list items *(i)* and *(ii)* due to the exclusion of large databases from the scope of data-driven modeling.

1.2 Data mining, data-driven modeling and wastewater treatment

Of the papers indexed by the ISI Web of Knowledge, a mere 0.05%[2] that deal with wastewater explicitly use the term "data-driven modeling" or "data mining" in their title, abstract or keywords. If only the papers that were published in the last five years are considered, the ratio is 0.1%[3]. Although this simple analysis neglects the publications that deal with a particular data mining or data-driven modeling technique without mentioning one of these two general terms, it still indicates, that the ratio of work that considers data mining and data-driven modeling techniques is relatively small. However, the growing number of publications in the last several years indicates a growing interest.

There are two explicit constraints for the methods developed and researched in this thesis; they define the niche to which the methods belong. First, the methods must target practical applicability. Second, they should neither require time-consuming nor costly measuring campaigns. The request for practical applicability is not restricted to the application of a given method on-site. Rather, adaptation to a specific plant and deployment on-site are equally important goals. Therefore, careful selection of the methods is crucial and consequently, aspects such as complexity, interpretability, robustness and reliability must be considered to allow safe use on-site. The avoidance of complex measuring campaigns is, from a practical point of view, essential. As recent publications state, because the possibility to carry out measuring campaigns on-site is very limited due to time and cost constraints, it is often necessary to rely on assumptions (Daigger, 2011; Dold *et al.*, 2010; Hauduc *et al.*, 2009).

[2]The query {Topic=("wastewater" AND ("data driven modeling" OR "data mining")), Timespan="All Years"} listed 51 entries, whereas the query {Topic=("wastewater"), Timespan="All Years"} listed 97,823 entries; queried on June 26th 2011.

[3]Same queries as above, however, the timespan was set to "Latest 5 years". The entry counts were 44 and 41,681, respectively; queried on June 26th 2011.

1.3 Goals and general research questions

Novel data mining and data-driven modeling techniques have great potential to support operators and engineers alike when applied to WWTP data. On the one hand, they can extract information valuable for plant operation and control from data available in the process information system of the plant. On the other hand, these methods can provide information from easy-to-measure data for which conventionally time-consuming experiments are required.

The following hypothesis can be formulated:

- WWTP data contain much more information valuable to support plant operation than is used today. Data mining and data-driven modeling methods can be used to extract this information.

The goal of this thesis is therefore to investigate the applicability and suitability of data mining and data-driven modeling techniques to support WWTP operation. Because the intended use of the resulting models is to apply them in practice, strategies must be developed to ascertain that the models can be adapted to new plants, that they can reliably be deployed and that their outputs can easily be interpreted.

The primary research questions that will be addressed in this thesis are:

- How can data mining and data-driven modeling techniques be applied to provide the operator information on past events (Chapter 3), to give him or her insights into the current state of the plant (Chapters 2 and 3) and to provide information helpful for the optimization of plant control (Chapters 2, 3 and 4)?
- How can the considered techniques be applied to provide the engineer more information on plant processes given the process data already available (Chapter 5), or by performing measuring campaigns that are easy to carry out (Chapters 4 and 5)?
- Can the complexity of the models be reduced in such a way that they can either be applied by non-experts or that non-experts can draw reliable conclusions from the results (Chapters 2 and 5)?
- To which extent can the modeling be automated and what is the role of available expert knowledge (Chapter 2 and 5)?
- Environmental conditions as well as WWTP processes are subject to change. How can these changes be detected, and how can they be considered (Chapters 2 and 3)?

1.4 Thesis outline

Chapter 2 Data-driven modeling approaches to support WWTP operation

Data-driven modeling techniques are evaluated for the set-up of software sensors. The methods rely on process data already available in the process information system and are tested considering three different levels of expert knowledge. It is shown that automatic software-sensor generation is possible. Because of the data-driven nature of the sensors, however, deployment must be performed carefully.

Chapter 3 Identification of industrial wastewater by clustering UV/Vis spectra

It is shown how information about industrial dischargers and their discharge patterns can be gained by clustering UV/Vis absorption spectra measured at the WWTP inlet. The cluster approach combines a self-organizing map with Ward clustering and could, in a full scale application, reliably detect the discharge events of an industrial laundry.

Chapter 4 Discharge distribution at hydraulic flow dividers

Neglecting the fact that most hydraulic flow dividers do not split the flows equally may have severe consequences not only for WWTP operation but also for modeling when even splitting is assumed. Up to now, there was no simple method available to check for and to quantify unequal distribution. In this chapter, a new method is introduced that applies dynamic time warping to compare the reactor influent and effluent signals of parallel reactors and thereby quantifies the discharge distribution among these reactors.

Chapter 5 Automatic reactor model synthesis with genetic programming

Successful WWTP modeling depends on the accurate description of the hydraulic processes. The accurate description, however, depends on information that cannot conveniently be obtained; hence, modeling is mostly based on experience and intuition. In this chapter, it is shown how genetic programming, a machine learning technique, can be used to synthesize reactor models based on already available or easy-to-measure data.

Please note that the simulation software for Chapters 3, 4 and 5 is available; see Appendix A on page 95.

References

Chan, C. W., 2003. Editorial: special issue on data-driven modelling methods and their applications. International Journal of Systems Science 34 (14), 731–732.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0: Step-by-Step Data Mining Guide.

Cios, K. J., Swiniarski, R. W., Pedrycz, W., Kurgan, L. A., Cios, K., Swiniarski, R., Kurgan, L., 2007a. The Knowledge Discovery Process. In: Data Mining. Springer, New York, NY, pp. 9–24.

Cios, K. J., Swiniarski, R. W., Pedrycz, W., Kurgan, L. A., Cios, K., Swiniarski, R., Pedrycz, W., Kurgan, L., 2007b. Introduction. In: Data Mining. Springer, New York, NY, pp. 3–7.

Daigger, G. T., 2011. A practitioner's perspective on the uses and future developments for wastewater treatment modelling. Water Science & Technology 63 (3), 516.

Dewasme, L., Bogaerts, P., Vande Wouwer, A., 2009. Monitoring of bioprocesses: Mechanistic and data-driven approaches. Studies in Computational Intelligence 218, 57–97.

Dold, P., Bye, C., Chapman, K., Brischke, K., White, C., Shaw, A., Barnard, J., Latimer, R., Pitt, P., Vale, P., Brian, K., 2010. Why Do We Model and How Should We Model? Proceedings of the 2nd Wastewater Treatment Modelling Seminar (WWTmod 2010), 133–149.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Magazine 17, 37–54.

Hauduc, H., Gillot, S., Rieger, L., Ohtsuki, T., Shaw, A., Takacs, I., Winkler, S., 2009. Activated sludge modelling in practice: an international survey. Water Science and Technology 60 (8), 1943–1951.

Maimon, O. Z., Rokach, L., 2005. Introduction to Knowledge Discovery in Databases. In: Maimon, O. Z., Rokach, L. (Eds.), Data mining and knowledge discovery handbook. Springer, Ramat-Aviv.

Solomatine, D., See, L. M., Abrahart, R. J., 2008. Data-Driven Modelling: Concepts, Approaches and Experiences. In: Abrahart, R. J., See, L. M., Solomatine, D. P. (Eds.), Practical Hydroinformatics. Vol. 68. Springer, Berlin, Heidelberg, pp. 17–30.

Solomatine, D. P., Ostfeld, A., 2008. Data-driven modelling: Some past experiences and new approaches. Journal of Hydroinformatics 10 (1), 3–22.

Verleysen, M., François, D., 2005. The curse of dimensionality in data mining and time series prediction. In: Computational Intelligence and Bioinspired Systems, Lecture Notes in Computer Science 3512. Springer, Berlin, Heidelberg, pp. 758–770.

Chapter 2

Data-driven modeling approaches to support wastewater treatment plant operation

Accepted for Publication in Environmental Modelling and Software

---

David J. Dürrenmatt and Willi Gujer

Data-driven modeling approaches to support wastewater treatment plant operation

David J. Dürrenmatt[a,b,*] *and Willi Gujer*[a,b]

[a] *Institute of Environmental Engineering, ETH Zurich, 8093 Zurich, Switzerland*
[b] *Swiss Federal Institute of Aquatic Science and Technology, Eawag, 8600 Dübendorf, Switzerland*
[*] *Corresponding author phone: +41 44 823 5407; fax: +41 44 823 5389; email: david.duerrenmatt@eawag.ch*

Abstract Data-driven modeling techniques are applied to process data from wastewater treatment plants to provide valuable additional information for optimal plant control. The application of data-driven modeling techniques, however, bears some risk because the generated models are of non-mechanistic nature and they thus do not always describe the plant processes appropriately. In this study, a procedure to build software sensors based on sensor data available in the process information system is defined and used to compare several techniques suitable for data-driven modeling, including generalized least squares regression, artificial neural networks, self-organizing maps and random forests. Three different degrees of expert knowledge are defined and considered mainly for optimum input signal selection and model interpretation. In two full-scale experiments, software sensors are created. The experiments reveal that even with linear modeling techniques, it is possible to automatically generate accurate software sensors. Hence, this justifies the selection of the most parsimonious and transparent models and to motivate their investigation by taking into account available expert knowledge. A high degree of expert knowledge is valuable for long-term accuracy, but can lead to performance decreases in short-term predictions. With regard to safe on-site deployment, the consideration of uncertainty measures is crucial to prevent misinterpretation of software-sensor outputs in the cases of rare events or model input failures.

Keywords Data-driven modeling; wastewater treatment plant; supervisory control and data acquisition; software sensor; generalized least squares; self-organizing map

2.1 Introduction

Increasing effluent quality requirements and fiscal restraints force wastewater treatment plant (WWTP) operators to fully exploit the available facilities. For optimal plant control, advanced control strategies must be implemented. These strategies require a plant-wide monitoring network of on- and off-line sensors, which are costly to acquire and maintain. It is in the operator's interest to both optimize the number of sensors installed and minimize the need for sensor maintenance. In this context, there is a need for cost-effective methods to provide the required information, ideally by directly extracting it from the data already

available in the supervisory control and data acquisition system (SCADA) and without the installation of additional instrumentation.

The deployment of software sensors is an alternative to the installation of hardware sensors (Cecil and Kozlowska, 2010). A software sensor is a piece of software that outputs a signal based on an internal model and several other (measured) input signals. Software sensors can be divided into three classes according to their underlying model, i.e., mechanistic, black-box and hybrid (gray-box) models (James, 2000). Because mechanistic models, such as the activated sludge model family (Henze *et al.*, 2000), often lack statistical identifiability, gray-box modeling approaches are preferred over mechanistic ones for the design of software sensors (Carstensen *et al.*, 1996). Gray-box modeling approaches include models with a reduced number of parameters that can be estimated by statistical or mathematical techniques. Numerous implementations based on state observers (e.g., Aubrun, 2001; Benazzi *et al.*, 2007; Lindberg and Carlsson, 1996; Sotomayor *et al.*, 2002), or stochastic models that incorporate process knowledge (e.g., Carstensen *et al.*, 1994) are available.

Software sensors based on black-box models are likewise popular because they do not require detailed prior understanding of the system (James, 2000). With data-driven modeling (DDM) techniques, it is cost-effective to set up software sensors based on black-box models. These techniques attempt to automatically capture the dominant processes and to relate input to output variables and can consequently be seen as an alternative when mechanistic models are not available or not valid (Gernaey *et al.*, 2004; Masson *et al.*, 1999). Popular techniques include multivariate statistical methods, such as multiple linear regression, principal component regression, partial least squares regression and artificial neural networks. Multivariate statistical methods can be used for, e.g., process monitoring (Lee *et al.*, 2006; Rosen *et al.*, 2003; Yoo *et al.*, 2004) and software-sensor design (Jansson *et al.*, 2002; Masson *et al.*, 1999), while artificial neural networks can be used for, e.g., prediction (Belanche *et al.*, 1999; Dellana and West, 2009; Kim *et al.*, 2009; Raduly *et al.*, 2007) and monitoring and control (Baeza *et al.*, 2002; Luccarini *et al.*, 2010).

From a practical point of view, successful deployment of a software sensor based on DDM depends on two critical aspects. First, to be competitive with hardware sensors, the setup must be fast and straightforward, while the quality of the sensor output must be comparable. In the optimal case, the sensor generation is widely automated, and the generated sensors rely only on data already available in the SCADA system. Second, for safe, long-term operation, it must be robust, i.e., it must reflect changing environmental conditions (e.g., changes in the wastewater composition, seasonal patterns) and changes in process control. This may be achieved with an adaptive underlying modeling technique or automatic periodical recalibration (Hill and Minsker, 2010; Rosen and Lennox, 2001), but at the very least, a sensor must provide a means of self-diagnosis to indicate critical changes. In addition, sensor input failures must be detected and isolated.

This paper investigates the applicability of four different modeling techniques for the derivation of cost-effective software sensors for use in a data-driven manner and with a strong focus on deployment. The techniques evaluated included generalized least squares regression (GLSR), artificial neural networks (ANNs), self-organizing maps (SOMs) and random forests (RFs). For sensor generation, three levels of expert knowledge are defined, ranging from the sparse knowledge of the types and locations of the sensors mounted in the WWTP to detailed knowledge of the hydraulic and biological processes. For comparability, a widely automated procedure is further introduced that allows the generation of software sensors based on available data and that takes into consideration the available knowledge.

The procedure is applied in two full-scale experiments carried out on a mid-sized nitrifying-denitrifying Swiss WWTP (65,000 people-equivalents). The only data source for the experiments is the SCADA system of the plant. In the first experiment, a software-sensor is generated to provide a redundant measurement of a hardware sensor measuring the concentration of the total chemical oxygen demand (COD) at the effluent of the primary clarifier. In the second experiment, a virtual on-online sensor that estimates the ammonia concentration in an activated sludge tank is developed for sensor diagnosis. The software sensors are compared in terms of accuracy and interpretability. The latter is important in this context because there is a risk that the models do not capture the dominant processes appropriately, which can only be detected in a transparent and interpretable model. Because the software sensors are targeted for on-site deployment, aspects such as robustness and long-term stability crucial to the operator will also be discussed.

2.2 Material and methods

2.2.1 Data-driven modeling

DDM is a legitimate alternative when mechanistic models cannot be applied. In DDM, data characterizing a system is analyzed to look for connections between the system state variables without taking into account explicit knowledge of the system's physical behavior. Methods used in DDM have developed from the fields of computational intelligence and machine learning (Solomatine *et al.*, 2008). Although DDM appears to be a cheap but limited technique, its simplicity of implementation, particularly when based on inexpensive, basic on-line measurement signals, is a compelling advantage (Dewasme *et al.*, 2009). Data-driven models are especially suitable whenever the rate of data acquisition surpasses the ability to analyze the data (Karim *et al.*, 2003), which is particularly true for WWTP operations. The main limitation of DDM is its inability to deal with changing conditions (e.g., modifications in process control, changing environmental conditions) when they are not included in the model and the lack of interpretability for some modeling techniques. If the data quality is poor or if there are no correlations between the response variable and the other variables, DDM will not be successful.

Because there is a risk that data-driven models do not capture the dominant processes appropriately, generated models must be carefully investigated, which requires expert-knowledge. The investigation, however, can be facilitated if transparent models are preferred and if the modeling is guided by the principle of parsimony, which recommends choosing the simplest possible explanation of a phenomenon (i.e. to prefer simple over complex models).

Selecting the most-suitable modeling technique for software-sensor generation is not trivial. Considering the dynamic nature of the WWTP processes, powerful non-linear modeling techniques may be preferred. However, taking into account the requirement that models, particularly those based on little expert knowledge, must be as transparent and interpretable as possible, parsimonious linear models may still be reasonable choices.

The best modeling techniques should not only predict a software-sensor value but also assign some measure of probability to it.

2.2.2 Procedure

A procedure was developed to systematically build software sensors, taking into account data available in the SCADA system. The procedure consists of the following six steps: *i)* requirement definition, *ii)* signal selection, *iii)* data preparation, *iv)* modeling, *v)* model evaluation and *vi)* deployment. This procedure will later be used to generate comparable software sensors based on the different levels of expert knowledge and using varying modeling techniques.

For this study, three different levels of expert knowledge are defined, as described in Table 2.1. Please note that the available knowledge is considered primarily for signal selection and model evaluation.

Step 1: Requirements definition

In this initial step, the requirements are formulated based on the planned use of the software sensor. This task is independent of the available expert knowledge and intends to define the desired properties of the software sensor, including the selection of the sensor to model, the type of analysis (on- or off-line), the required accuracy and the desired sampling rate. The latter is particularly dependent on the intended use. The upper bound is set by the sampling rate of the stored signals and the lower bound by the processes of interest. While signals with low sampling rates may exhibit less noise, they may not be suitable for plant control anymore.

Step 2: Signal selection

Signal selection depends on the desired sampling rate of the software sensor as well as on the level of expert knowledge available and the signals stored in the SCADA system. For the derivation of an on-line software sensor, only the hardware sensors located upstream of the software-sensor location are meaningful input signals unless there are recycled flows, or if the sensor measures a quantity that propagates fast (e.g., wastewater discharge). Off-line sensors, i.e., those used for retrospective analyses, can include both downstream and upstream sensors. Because even the Basic Knowledge scenario includes information about the sensor locations, signal selection can always be made. In addition, knowledge of the sensor types can be used to exclude knowingly unreliable input sensors.

Table 2.1: Characteristics of three levels of expert knowledge a priori available for software-sensor development.

Type	Knowledge
Basic Knowledge (BK)	Types and locations of existing sensors
Intermediate Knowledge (IK)	Basic Knowledge plus approximate hydraulic delays between the sensors
Advanced Knowledge (AK)	Intermediate Knowledge plus detailed information on reactor hydraulics and biological processes

Step 3: Data preparation

All selected signals are first preprocessed by the robust regression-based trimmed repeated median (TRM) filter for outlier removal (Bernholt *et al.*, 2006). The window width, which should be at least three times the length of the outlier patches, was fixed in this study at 31 time steps.

Afterwards, all variables in the dataset are down-sampled to the desired sampling interval Δt.

Lagged variables are then introduced to reflect values at the present time t of previous times, $t - \alpha\Delta t$, with lag α. Lagging is applied to account for the hydraulic delays between the locations of the software sensor and each input variable and to take into account the dynamics of the process (Masson *et al.*, 1999). For the generation of on-line sensors, only positive lags ($\alpha \geq 0$) are considered. We suggest adding lagged variables rather generously if the available knowledge is limited. However, if Intermediate Knowledge or Advanced Knowledge is available, the lagging should be limited to specific, physically meaningful lags.

Log-transformed versions of the variables (including the lagged variables) are appended to the dataset. The motivation of the log transformation is threefold: *i)* to convert between multiplicative and additive relationships, i.e., linearization (e.g., $\log(QC) = \log(Q) + \log(C)$); *ii)* to alleviate possible issues due to heteroscedasticity, i.e., to stabilize the variance in the residuals and *iii)* to dampen exponential growth patterns (e.g., $\log(a^b) = b\log(a)$). Other transformations that would also be conducive to modeling are abandoned for the following several reasons: each transform doubles the number of variables in the data set, they are sometimes difficult to interpret and some can be automatically built during modeling anyway (e.g., n-th order differences in a linear model by subtracting lagged variables).

The nonspecific inclusion of input sensor candidates, the large number of lags and the applied transformations lead to high-dimensional data sets with highly correlated variables, which can be challenging for many methods (Rosen and Lennox, 2001). A method often applied to reduce the dimensionality is principal components analysis (PCA) (Hastie *et al.*, 2009; Montgomery *et al.*, 2006). PCA projects the data space onto a space spanned by principal components, which are linear combinations of the variables of the data space and are chosen as follows. The first component has the direction of the largest variance. The next components are chosen orthogonal to the previous components in the direction of the largest remaining variance. The dimensionality is then reduced by only considering the components that describe a defined amount of the total variance. Due to the orthogonality of the principal components, they are uncorrelated.

Depending on the type of analysis performed, the data set is divided. To assess the prediction error of a particular model, split-sample validation is used. That is, the data are divided into a training set to fit the model and a validation set to calculate the error. On the other hand, to compare several modeling techniques, ten-fold cross-validation is applied to compute the generalization error (Hastie *et al.*, 2009). For this purpose, the dataset is divided into ten equal-sized subsets. Ten models are built, each time using nine subsets for calibration and leaving out a different subset to assess the error, i.e. to calculate the residuals. The generalization error is then calculated given the residuals of all models.

Step 4: Modeling

In this step, a model is created with the training set. The modeling techniques considered in this paper are the generalized least squares regression (GLSR), artificial neural networks (ANN), self-organizing maps (SOM) and random forests (RF), which will be introduced in Section 2.2.3.

Step 5: Model evaluation

The calibrated model is evaluated with the validation set and it is tested to determine whether the defined performance criteria are met. If this is the case, the model is accepted and can be deployed. If not, the previous steps must be repeated. The performance criterion used in this paper is the coefficient of variation of the root mean square deviation, CV(RMSD), which is calculated by

$$CV(RMSD) = \frac{1}{\overline{y_{obs}}}\sqrt{\frac{\sum_{i=1}^{N}\left(y_{obs,i} - y_{mod,i}\right)^2}{N}} \tag{2.1}$$

where y_{obs} denotes the series of observations and y_{mod} the series of model predictions for both vectors with length N.

Step 6: Deployment

The evaluated model can now be deployed as a software sensor to assist WWTP operation. Software-sensor values are available at the chosen sampling rate. Because of the non-intelligent data-based nature of the software sensor, it may not automatically adapt to changes in the process and changes in the catchment, and it may not be able to cope with rare events. Therefore, supervision of the software sensor during deployment is important, and some means for self-diagnosis, depending on the modeling technique, must be made available to users (Masson *et al.*, 1999).

2.2.3 Modeling techniques

In the section, the considered modeling techniques are briefly introduced. Additional information on the techniques, including additional references, can be found in the Supporting Information.

Generalized least squares regression (GLSR)

Generalized least squares regression is a linear modeling technique. In contrast with the ordinary least squares (OLS) estimation method, GLS estimation does not rely on the assumption that the residuals are uncorrelated and have constant variance. This is an important feature because auto-correlated residuals typically occur when dealing with time series data and incomplete models (Dellana and West, 2009).

Backward-elimination is applied to find the best subset of regressor variables by taking into account the BIC (Bayesian information criterion), which not only considers the quality of the fit, but also penalizes complex models (Hastie *et al.*, 2009).

Artificial neural network (ANN)

The artificial neural network is a popular supervised non-linear statistical data modeling tool. In this paper, a multilayer perceptron with one hidden layer is considered, which is a feed-forward ANN. All neurons are given a sigmoidal activation function, except the neurons in the output layer, which have a linear activation function. The network is trained with a back-propagation learning (Hastie *et al.*, 2009) and early-stopping is applied to prevent over-fitting. The optimal number of hidden units is problem-dependent and therefore assessed for each data set by training several networks with a varying number of hidden neurons, eventually selecting the network with lowest generalization error.

Self-organizing maps (SOM)

Self-organizing maps are a noise-tolerant variant of artificial neural networks based on unsupervised learning, originally proposed by Kohonen (2001). They learn to project input data in a non-linear fashion from a high-dimensional data-space onto a lower-dimensional discrete lattice of neurons on an output layer, called feature map (Céréghino and Park, 2009; Kalteh *et al.*, 2008). This is done in a topology-preserving way, which means that neurons physically located close to each other have similar input patterns.

Each neuron has assigned a prototype vector having the same dimensionality as the input data. The quantification error, *q.e.*, expresses how well an input vector is represented in the SOM and can be considered as a means for software-sensor self-diagnosis (i.e., the higher the *q.e.*, the more uncertain the prediction).

The models presented in this paper all have a two-dimensional, hexagonal feature map. The number of neurons and the ratio of the side lengths are determined taking into consideration the size of the data set (Park *et al.*, 2006; Vesanto and Alhoniemi, 2000). The measure of topological relevance (MTR) is considered to rank the importance of the variables (Corona *et al.*, 2008).

Random forests (RF)

Random forest is an increasingly popular machine-learning technique (Mouton *et al.*, 2011; Verikas *et al.*, 2011). RFs are non-linear ensemble classifiers that build on a large collection of classification or regression trees that are aggregated (Breiman, 2001). The RF technique has the advantage that it performs remarkably well with very little tuning required (Hastie *et al.*, 2009) and is not prone to over-fitting (Breiman, 2001; Hastie *et al.*, 2009); hence, it is suitable for highly automated data-driven modeling approaches.

If RF is applied for regression, the response of the RF is the averaged response of all trees. The relative importance of the regressor variables can be measured with samples not selected in the bootstrap sub-samples used to construct a tree (Hastie *et al.*, 2009; Verikas *et al.*, 2011).

2.3 Results and discussion

2.3.1 Full-scale experiments

The given procedure was applied to create software sensors in two full-scale experiments. The modeling technique as well as the degree of expert knowledge available was varied to assess the suitability of each method and the role of expert knowledge. The experiments were carried out on a nitrifying-denitrifying WWTP which treats the wastewater of 65,000 people-equivalents. It is composed of two activated sludge treatment stages (cf. Figure 2.1). The first, older, fully aerated "BB" stage consists of two parallel lanes and is used to pre-treat approximately 50% of the wastewater leaving the primary clarifier. The newer "NB" stage consists of four lanes operated in parallel, with the first zone of each reactor operated anoxically or oxically, depending on the ammonia load. The only data source for the experiments was the SCADA system of the plant; all signal data were stored with a sampling interval of three minutes.

In the first experiment, an on-line software sensor was designed to provide a redundant measurement of a hardware sensor measuring the concentration of the total COD at the effluent of the primary clarifier, C_{COD}. The hardware sensor, a submersible spectrometer probe (spectro::lyzer by s::can Messtechnik GmbH, Vienna, Austria) with 5 mm optical path length, estimates the COD based the absorption at different wavelengths using an internal model calibrated to the local situation (cf. Langergraber *et al.*, 2003). The availability of a redundant measurement that can be used for sensor diagnosis is valuable because the hardware sensor is mounted in heavily polluted wastewater and therefore subject to fouling. In the optimal case, the software sensor would replace the hardware sensor. For this purpose, it must be reliable and exhibit long-term accuracy. A sampling interval of 15 minutes was chosen, which reduces noise while still enabling the use of the sensor for process control and trend detection.

The signals selected for this experiment are listed in Table 2.2. The accuracy of the software sensors generated with different modeling techniques and under consideration of a varying degree of expert knowledge is given in Table 2.3. The accuracy was estimated with ten-fold cross-validation using a 21-day data set (1st– 21^{st}October 2009).

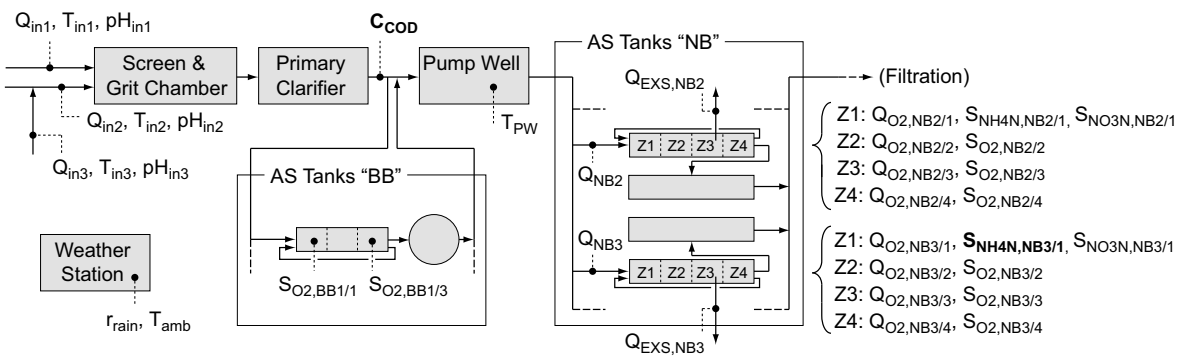

Figure 2.1: Simplified layout of the nitrifying-denitrifying WWTP. The locations of the signals selected for the experiments are indicated, the sensors modeled by software sensors are marked in boldface (C_{COD} for the first and $S_{\mathrm{NH4N,NB3/1}}$ for the second experiment). Q_{O2} relates to airflow.

Table 2.2: Signals selected for both full-scale experiments. For each degree of expert knowledge (BK = Basic Knowledge, IK = Intermediate Knowledge, AK = Advanced Knowledge), it is indicated whether a sensor was considered ("Use") and which lags were added to the dataset ("Lags"). For every sensor the log transform was calculated. The sensor locations are given in Figure 2.1. Mean and standard deviation were calculated for the periods given in the text. (Q: discharge, S: concentration of soluble matter, C: concentration of soluble and particulate matter, T: temperature; Q_{exs}: excess sludge removal, r_{rain}: rainfall intensity).

			BK		IK		AK	
Signal	Unit	Mean ± Std. Dev.	Use	Lags	Use	Lags	Use	Lags
Experiment 1 (Response: C_{COD}, 530 ± 150 g/m^3)								
Q_{in1}	m^{3}3/s	65 ± 39	X	0-10	X	0-10	X	0, 1
pH_{in1}	–	8.3 ± 0.16	X	0-10	X	0-10		
T_{in1}	°C	19 ± 1.1	X	0-10	X	0-10		
Q_{in2}	m^3/s	72 ± 39	X	0-10	X	0-10	X	0, 1
pH_{in2}	–	7.8 ± 0.18	X	0-10	X	0-10		
T_{in2}	°C	20 ± 1.0	X	0-10	X	0-10		
Q_{in3}	m^3/s	56 ± 29	X	0-10	X	0-10	X	0, 1
pH_{in3}	–	7.5 ± 1.1	X	0-10	X	0-10		
T_{in3}	°C	19 ± 2.6	X	0-10	X	0-10		
T_{PW}	°C	20 ± 1.1	X	0-10	X	0		
$S_{O2,BB1/1}$	g/m^3	1.2 ± 0.85	X	0-10	X	0, 1, 2	X	0, 1, 2
$S_{O2,BB1/3}$	g/m^3	2.4 ± 2.1	X	0-10	X	0, 1, 2	X	0, 1, 2
$S_{NH4N,NB2/1}$	g/m^3	10.2 ± 2.5	X	0-10	X	0, 1, 2	X	0, 1, 2
T_{amb}	°C	12 ± 5.3	X	0-10	X	0-10		
Experiment 2 (Response: $S_{NH4N,NB3/1}$, 8.7 ± 2.6 g/m^3)								
Q_{tot} [a]	m3/s	180 ± 120	X	0-10	X	0, 1, 2	X	0, 1, 2
T_{PW}	°C	16 ± 0.91	X	0-10				
Q_{NB2}	m^3/s	41 ± 33	X	0-10	X	0, 1, 2	X	0, 1, 2
$Q_{EXS,NB2}$	m^3/hr	0.21 ± 0.12	X	0-10	X	0-10	X	0-10
$Q_{O2,NB2/1}$	m^3/hr	130 ± 190	X	0-10	X	0, 1	X	0, 1
$S_{NH4N,NB2/1}$	g/m^3	12 ± 7.0	X	0-10	X	0	X	0
$S_{NO3N,NB2/1}$	g/m^3	6.0 ± 5.0	X	0-10	X	0	X	0
$Q_{O2,NB2/2}$	m^3/hr	470 ± 210	X	0-10	X	0, 1	X	0, 1
$S_{O2,NB2/2}$	g/m^3	1.3 ± 0.60	X	0-10	X	0, 1	X	0, 1
$Q_{O2,NB2/3}$	m^3/hr	410 ± 170	X	0-10				
$S_{O2,NB2/3}$	g/m^3	1.8 ± 0.93	X	0-10				
$Q_{O2,NB2/4}$	m^3/hr	310 ± 180	X	0-10	X	0, 1	X	0, 1
$S_{O2,NB2/4}$	g/m^3	2.8 ± 1.6	X	0-10	X	0, 1	X	0, 1
Q_{NB3}	m^3/s	55 ± 28	X	0-10	X	0, 1, 2	X	0, 1, 2
$Q_{EXS,NB3}$	m^3/hr	0.31 ± 0.10	X	0-10	X	0-10	X	0-10
$Q_{O2,NB3/1}$	m^3/hr	16 ± 82	X	0-10	X	0, 1	X	0, 1
$S_{NO3N,NB3/1}$	g/m^3	4.7 ± 6.1	X	0-10	X	0	X	0
$Q_{O2,NB3/2}$	m^3/hr	580 ± 210	X	0-10	X	0, 1	X	0, 1
$S_{O2,NB3/2}$	g/m^3	1.1 ± 0.51	X	0-10	X	0, 1	X	0, 1
$Q_{O2,NB3/3}$	m^3/hr	470 ± 170	X	0-10				
$S_{O2,NB3/3}$	g/m^3	1.8 ± 0.81	X	0-10				
$Q_{O2,NB3/4}$	m^3/hr	290 ± 120	X	0-10	X	0, 1	X	0, 1
$S_{O2,NB3/4}$	g/m^3	2.31 ± 0.83	X	0-10	X	0, 1	X	0, 1
T_{amb}	°C	14 ± 4.4	X	X 0-10		X 0		
r_{rain}	mm/min	0.0023 ± 0.010	X	0-10	X	0-10	X	0-10

[a] $Q_{tot} = Q_{in1} + Q_{in2}$

Table 2.3: Comparison of the performance achieved by the different modeling techniques with varying levels of expert knowledge. The performance is expressed as $CV(RMSD)$, cf. Eq. (2.1) and evaluated with ten-fold cross-validation over 21 days. (BK = Basic Knowledge, IK = Intermediate Knowledge, AK = Advanced Knowledge; see Table 2.1).

Modeling Technique	Dimensionality Reduction	Experiment 1: C_{COD}			Experiment 2: $S_{NH4N,NB3/1}$		
		BK	IK	AK	BK	IK	AK
GLSR	None	0.12	0.13	0.19	0.10	0.10	0.11
GLSR	PCA	0.15	0.16	0.22	0.11	0.12	0.12
ANN[a]	None	0.03	0.03	0.11	0.06	0.08	0.10
ANN [a]	PCA	0.03	0.03	0.14	0.08	0.09	0.10
SOM	None	0.06	0.06	0.16	0.08	0.09	0.09
SOM	PCA	0.07	0.07	0.16	0.14	0.15	0.16
RF	None	0.04	0.04	0.09	0.06	0.06	0.07
RF	PCA	0.04	0.05	0.11	0.06	0.08	0.08

[a] Number of hidden neurons. Exp 1: a) BK: 59, IK: 62, AK: 81, b) BK: 55, IK: 57, AK: 80; Exp 2: a) BK: 25, IK: 55, AK: 38, b) BK: 9, IK: 11, AK: 16

In the second experiment, a virtual on-line sensor that estimates the NH_4-N concentration in the anoxic zone of an activated sludge tank (AST) was designed with the purpose of sensor diagnosis. The sensors to be diagnosed are in-situ ion-selective electrodes (Nadler Chemische Analysetechnik AG, Zuzwil, Switzerland); their membranes have a lifetime of three months in activated sludge (for further information, cf. Rieger *et al.*, 2002). In two of the four ASTs operated in parallel, a sensor is installed and considered for the control of the aeration in the first zone of two ASTs (sensor in NB2 for NB1 and NB2, sensor in NB3 for NB3 and NB4; cf. Figure 2.1). The zones are anoxically operated, unless the ammonium load surpasses a certain limit, which causes continuous aeration to assure sufficient nitrification. Sensor failures can have a fatal impact on overall plant performance. Therefore, careful maintenance and monitoring are important. For brevity, only the generation of a redundant sensor for $S_{NH4N,NB3/1}$ mounted in NB3 is discussed.

The software sensor designed for sensor diagnosis must be accurate but does not necessarily need to have a long lifetime because frequent recalibration is possible. It should have a sampling interval of 15 minutes.

The signals selected for sensor generation are indicated in Figure 2.1 and described in Table 2.2. The accuracy of the resulting sensors based on different modeling techniques and levels of expert knowledge, with and without dimensionality reduction, is given in Table 2.3. They were calculated with ten-fold cross-validation using a 21-day data set (15^{th}April – 5^{th}May 2010).

2.3.2 Comparison of the modeling techniques

Performance

The results in Table 2.3 show that for both experiments, models can be set up that are able to reproduce the behavior of the selected signal.

Generally, the performance of the non-linear models (ANN, SOM and RF) is superior to the performance of the linear GLSR model, independently of the knowledge available. This is not

surprising because both hydraulic and biological processes are known to be highly non-linear and dynamic (Belanche *et al.*, 1999; Dellana and West, 2009).

Models applying PCA for dimensionality reduction perform worse than those without. This is due to the information loss caused by only considering the principle components that describe 95% of the variance of the regressor variables. This is unaffected by the response variable (Jackson, 1991). The effect of PCA seems particularly severe when considering the SOM models. This, however, is partially attributed to another effect. Because the size of the SOM has been made dependent on shape of the data set, significantly smaller maps with limited modeling capabilities result for the data sets reduced in dimensionality.

Increasing expert knowledge is correlated with decreasing accuracy. Higher levels of expert knowledge lead to smaller data sets with more physically sound variables, but fewer variables which might show some local correlations and lead to better model accuracy. As will be shown later, not considering these correlations is beneficial for the application in the long-run.

Transparency

Because there is a risk that data-driven models do not capture the dominant processes appropriately, the models should be carefully checked. Model checking requires transparency of the model and expert knowledge to distinguish a meaningful from a not-so-meaningful model. However, even when expert knowledge is lacking, good interpretability is important because it adds to the available knowledge.

Model transparency will be discussed using the models of Experiment 1 with Advanced Knowledge.

The GLSR models are highly transparent; their linear model can be expressed as a single equation. The formula for the GLSR model (without dimensionality reduction) is

$$\begin{aligned} C_{\mathrm{COD}} \;=\; & 5Q_{\mathrm{in1}} - 6\ \mathrm{lag}_1(Q_{\mathrm{in1}}) - 417\log(Q_{\mathrm{in1}}) + 403\ \mathrm{lag}_1\left(\log\left(Q_{\mathrm{in1}}\right)\right) \\ & +3Q_{\mathrm{in2}} - 464\log(Q_{\mathrm{in2}}) + 537\ \mathrm{lag}_1(\log(Q_{\mathrm{in2}})) \\ & -18S_{\mathrm{O2,BB1/3}} + 16\ \mathrm{lag}_2(S_{\mathrm{O2,BB1/3}}) \\ & -814\log(S_{\mathrm{NH4N,NB1/1}}) + 808\ \mathrm{lag}_2(\log(S_{\mathrm{NH4N,NB1/1}})) + 217 \end{aligned} \tag{2.2}$$

where the functions $\mathrm{lag}_\alpha(\cdot)$ denote the signal shifted in the time domain by α time steps. Of many variables, there are differently lagged pairs with similar coefficients subtracted from each other in the formula. This indicates that the gradient, i.e., the change of a variable over time, is important rather than the absolute value. If the export had only added one single lag of the given variables, however, it would not have been possible to take into account these gradients. The significance of the selected variables evaluated with an F-test reveals that all variables are significant at the 99% significance level. The interpretation of the regression coefficients, however, can be misleading because they might be skewed by collinearity. Although one sees that, roughly, high C_{COD} corresponds with low Q_{in1}, high Q_{in2}, low $S_{\mathrm{O2,BB1/3}}$ and low $S\mathrm{NH4N,NB1/1}$, further interpretation would require more information on the catchment.

Applying PCA for dimensionality reduction prevents collinearity issues because the principal components are orthogonal, but then the interpretation of the original model variables and their importance are less obvious (Jackson, 1991).

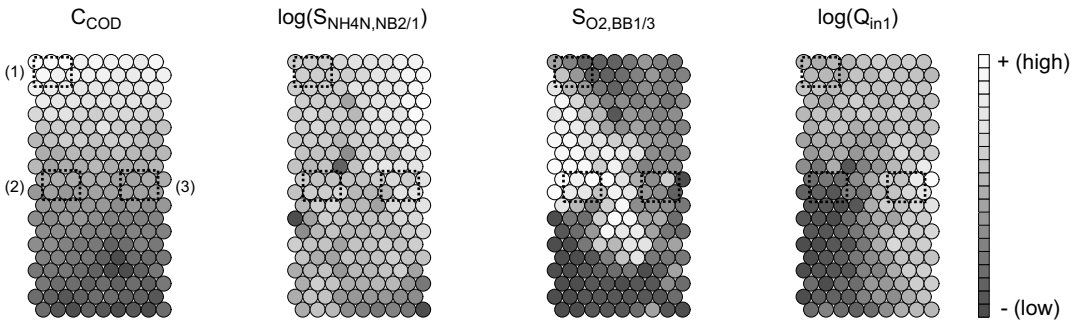

Figure 2.2: SOM content planes of C_{COD} (response) and the three regressor variables with the highest topological relevance (MTR). Properties of the regions (1)-(3) are discussed in the text.

The ANN models have better performance but, in contrast, cannot easily be interpreted, they are essentially opaque. This is also the primary reason for which they are criticized (Paliwal and Kumar, 2011) and why their use is limited in fields where model interpretation is important (Hastie *et al.*, 2009).

The non-linear mappings of the SOM can conveniently be assessed by plotting the content planes, which are basically cross-sections through the prototype vectors of the feature map. The planes of the three variables with the highest MTR are plotted together with C_{COD} in Figure 2.2, three regions are marked. The analysis indicates that the highest values of C_{COD} occur at rather high $S_{\mathrm{NH4N,NB2/1}}$ and medium to low $S_{\mathrm{O2,BB1/3}}$ (region 1). This corresponds to the interpretation of the GLSR model, given in Eq. (2.2). However, medium C_{COD} values can occur at high $S_{\mathrm{O2,BB1/3}}$ with low Q_{in1} and vice versa (regions 2 and 3).

If PCA is applied to reduce the dimensionality of the SOM, however, the advantage of the high interpretability is lost.

The performance of the models generated with RF without dimensionality reduction is comparable to the ANN. The interpretability, however, is likewise limited (Mouton *et al.*, 2011). The variable importance measure reveals that the most important variables are Q_{in1}, $S_{\mathrm{O2,BB1/3}}$ and $S_{\mathrm{NH4N,NB2/1}}$ (in decreasing order) and thereby suggests the same variables as GLSR and SOM. Similar to the other modeling techniques, applying PCA dimensionality reduction worsens interpretability of even the variable importance measures.

Even though the modeling techniques are very different, the generated models consistently consider the same input signals important. With regard to the interpretability, however, only the GLSR and SOM models without dimensionality reduction by PCA are sufficiently interpretable.

2.3.3 Role of expert knowledge

The role of expert knowledge is not directly evident from the performance measures given in Table 2.3. On the contrary, it seems that a higher degree of knowledge leads to lower accuracy. The role of expert knowledge becomes clear when the models are applied for longer term predictions. In Figure 2.3, the cumulated absolute residuals for a long-term prediction of C_{COD} with the GLSR are plotted. A calibration period of seven days was used to calibrate

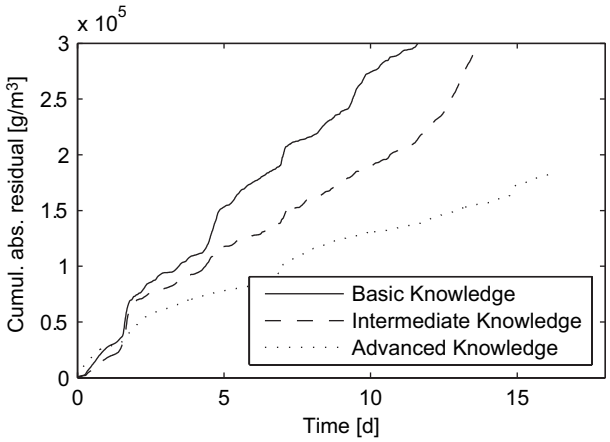

Figure 2.3: Cumulative absolute residuals for a long-term prediction of C_{COD} using a GLSR model.

the models. During the first 1.5 days of the prediction, the increase is linear and similar for all of the levels of knowledge. A linear increase means that the residuals remain approximately constant for the time considered. After this, while the cumulative absolute residual of the sensor based on Advanced Knowledge still increases almost linearly, the cumulated absolute residuals of the others exhibit faster increases. This indicates that models based on a higher degree of knowledge give more consideration to the true physical processes and have a lower tendency to over-fit local correlations that are valid for short time spans only. The sensor based on Advanced Knowledge is given in Eq. (2.2).

As a result, long-term accuracy is sensitive to the available degree of expert knowledge. Expert knowledge is therefore important if long-term accuracy is a key goal and if frequent recalibration is not possible.

2.3.4 Deployment aspects

For clarity and brevity, the deployment aspects are discussed considering the two modeling techniques with the highest interpretability only, namely GLSR and SOM without dimensionality reduction by PCA.

Long-term accuracy

Long-term accuracy is an important feature to ensure cost-effectiveness, particularly if the setup of the software sensor has required additional measuring campaigns, and also with regard to reliable application. However, the automatically generated software sensors based on empirical models will only have a limited lifetime. Some measures, such as prediction intervals or quantification errors, can be applied as indicators for decreasing accuracy, but comparing grab samples with the software-sensor outputs is also appropriate for software-sensor diagnosis. The latter is a common task for sensor diagnosis and fault detection, and highly optimized procedures are available (e.g., Corominas *et al.*, 2011; Rieger *et al.*, 2004).

In Section 2.3.3, it was shown that a high degree of expert knowledge can have a positive effect on long-term stability because the expert excludes sensors known to be either inaccurate or

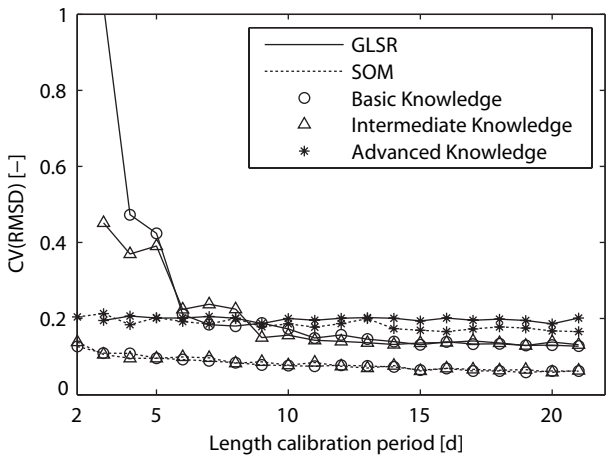

Figure 2.4: $CV(RMSD)$ of GLSR and SOM models (without PCA) for Experiment 1. The models were calibrated with variable length calibration sets and validated with a one-day validation set.

to obey a trend. Trends, such as weekly and seasonal patterns, however, are not an issue if the software sensor is designed to have a lifetime shorter than that of the respective pattern or if the software sensor is adaptive (Rosen and Lennox, 2001). If long-term applicability is a principal goal, the input data must be checked for trends and non-stationarity, e.g., by visual inspection and statistical tests (Montgomery *et al.*, 2006), for which a high degree of expert knowledge is helpful.

The length of the time series used to calibrate the sensors has an impact on their stability and can, to some extent, substitute for expert knowledge. Short calibration periods tend not to properly catch the relevant processes, as shown in Figure 2.4. In a cross-validation process, the $CV(RMSD)$ was calculated for the GLSR and SOM models built for Experiment 1. The calibration sets had varying lengths, while the length of the validation period was set to one day.

For GLSR, models calibrated with short calibration periods and with Basic and Intermediate Knowledge result in high deviations. This indicates not only that the relevant processes cannot be identified with short calibration periods, but also that the model is over-fitted. The $CV(RMSD)$ converges after a calibration set length of approximately 10 days. Only the model with Advanced Knowledge performs equally well for any investigated length of the calibration period. Similar effects can be identified for the SOM model. However, severe over-fitting is prevented by the connection of the SOM-map size and the size of the calibration set (the smaller the set, the smaller the map). The $CV(RMSD)$ of the SOM model is generally lower than that of the GLSR model because of its ability to model non-linear processes to a greater extent. For Advanced Knowledge, however, a significantly higher $CV(RMSD)$ is observed, which indicates that some sensors and lags were removed from the data set that would have been potentially useful.

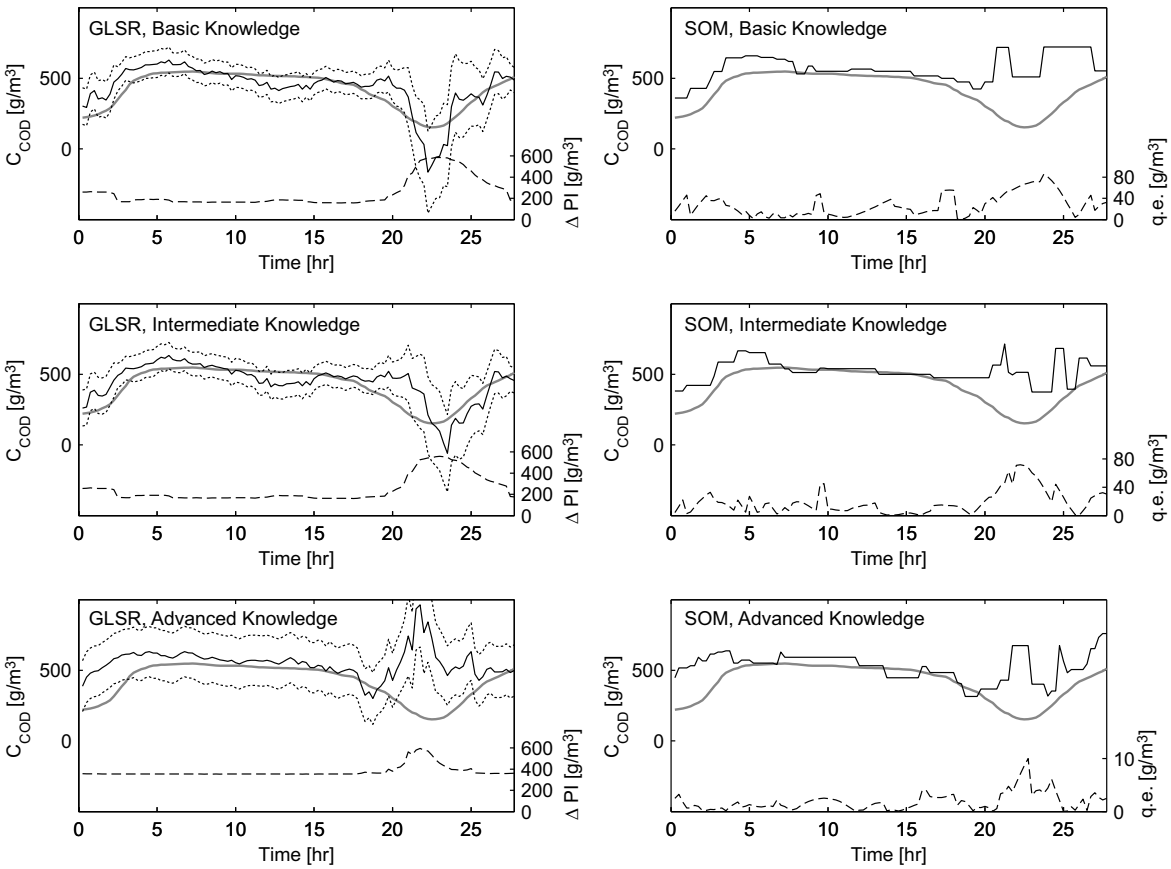

Figure 2.5: C_{COD} Software sensor responses to a rainfall event. The software sensors are based on the GLSR (left column) and SOM models (right column). The measured (bold grey line) and calculated (solid black line) C_{COD} are plotted. For the GLSR, the 95% prediction interval (dotted line) and the ΔPI (dashed line, right axis) are indicated, and for the SOM, the quantification error, *q.e.* (dashed line, right axis) is plotted.

Rare events

To be of use to the operator, rare events should either be correctly modeled by the software sensor, or the software sensor should indicate that some unknown event has happened. The output of the software sensors with a GLSR or SOM model for Experiment 1, both without dimensionality reduction by PCA, were tested against a rainfall event. As indicators for the probability of a prediction, the 95% prediction intervals, PI, and the difference between the upper and the lower PI, ΔPI, were calculated for the GLSR model. For the SOM model the quantification error, *q.e.*, was calculated considering 33% of the most important variables (ranked by the MTR measure). The results are plotted in Figure 2.5.

During the rainfall event (the influent discharge rose from $\sim 0.1\mathrm{m}^3/\mathrm{s}$ to $\sim 0.6\mathrm{m}^3/\mathrm{s}$ from hour 20 to hour 22), both ΔPI and *q.e.* markedly increase and thereby reliably signal the rare event. The software-sensor responses during the event can be classified into two groups. The sensors that correlate high C_{COD} values to high temperature under-estimate C_{COD} due to the rather cold temperature of the rainfall. This is the case for the GLSR models based on Basic and Intermediate Knowledge. The other models correlate high C_{COD} to high influent discharge and over-estimate C_{COD}.

The monitoring of ΔPI and $q.e.$ is a valuable tool and essential for safe software-sensor deployment.

Model transferability

The purpose of Experiment 2 is to generate a software sensor to predict the NH_4-N concentration in lane 3 (Figure 2.1) based on other measurements for sensor diagnosis. It is a legitimate question whether it would be possible to calibrate a model for lane 2 because there is a NH_4-N probe permanently mounted and to transfer that model to the other lanes.

The transferability was assessed as follows, with a data set of one year and by taking into account Advanced Knowledge. First, a model was calibrated for lane 2 using data from days 1-14 (all signals except the measurements in the other lanes), and the generalization error was calculated using data from day 15. Then the model was transferred to lane 3, i.e., all signals measured in lane 2 were replaced by the corresponding signal in lane 3, and the error was calculated using the data from day 15. Another model was then calibrated with data from days 2-15 and validated with data from day 16. This was repeated until day 365 was used for validation. The average $CV(RMSD)$s are given in Table 2.4.

The performance of the transferred models is worse than the performance of the original model, independent of the applied modeling technique. The reasons for this discrepancy can be manifold, e.g., unequal distribution of the wastewater streams or different aerator efficiencies. The decision whether the transferability is enough, however, depends on the required accuracy of the sensor and therefore cannot be generally answered here.

Model input failure

Until now it was assumed that the software-sensor inputs were not subject to failure. Although the failure rate may be considerably reduced if robust sensors are selected as inputs, there must be a way to identify potential failures. This is of particular importance if the software sensor is used for sensor diagnosis because failures of sensors to diagnose must be distinguished from failures of sensors considered as software-sensor inputs.

A straight forward procedure to identify whether an input sensor has failed is to create a primary software sensor using the full data set and to create secondary software sensors using reduced data sets. That is, each reduced data set contains the signal data of all but one sensor (including its transformed and lagged variables). During deployment, if the primary software sensor changes behavior, the secondary sensors can be checked. If all but one sensor show similar changes, the signal left out for that individual secondary sensor is probably the

Table 2.4: Assessment of the model transferability. A GLSR model and a SOM model were calibrated for lane 2 and then transferred to lane 3. The generalization error expressed as $CV(RMSD)$ is given for lane 2 and lane 3.

Lane	GLSR	SOM
Lane 2 (Original model)	0.16	0.17
Lane 3 (Transferred model)	0.31	0.28

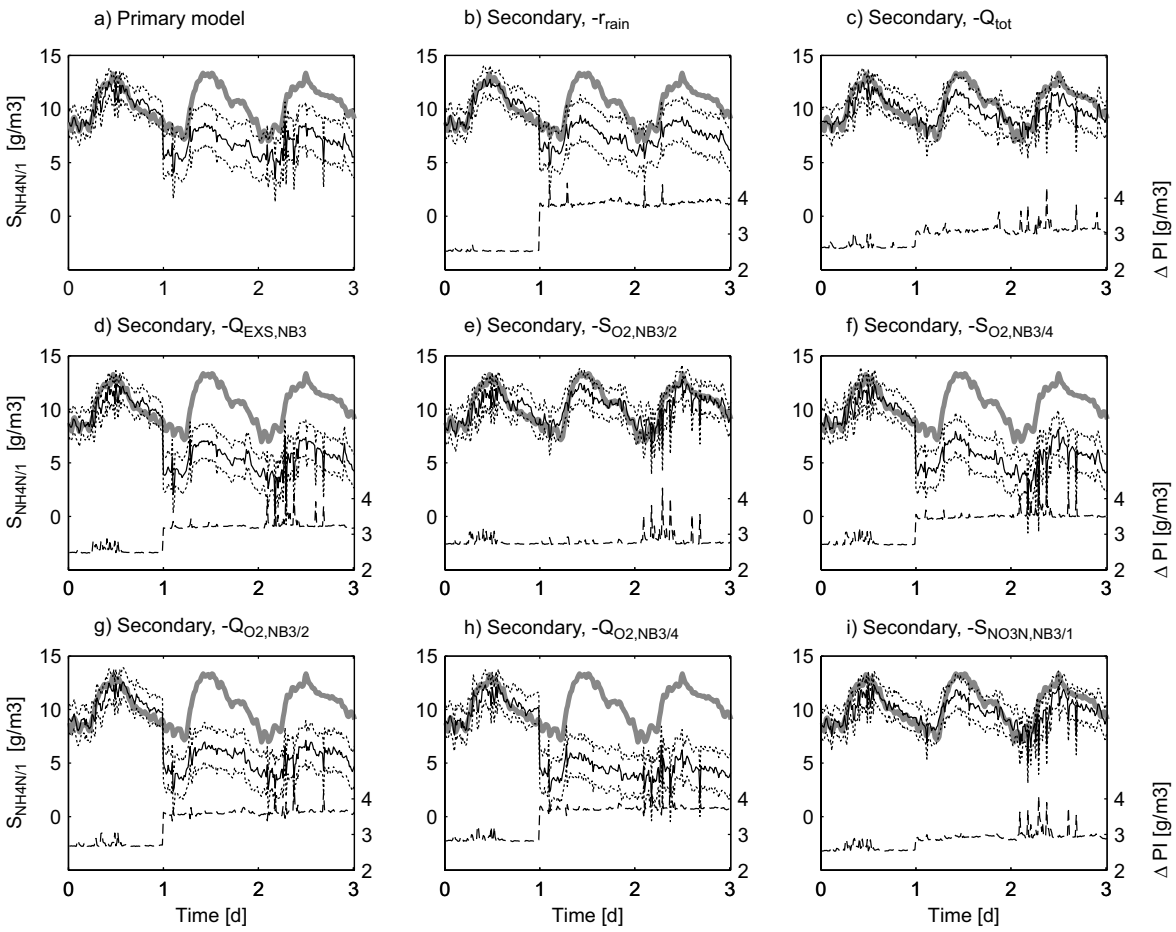

Figure 2.6: Results of the validation period for a software sensor to model $S_{\mathrm{NH4N,NB3/1}}$. The input sensor $S_{\mathrm{O2,NB3/2}}$ fails after one day. In subplot a), the measured NH_4-N concentration (thick grey line) and the output of the primary model based on the full data set (solid line for the prediction, dotted lines for the 95% prediction intervals) are shown. In subplots b) to i), the outputs of the secondary models based on reduced data sets are shown, ΔPI is indicated (dashed lines).

one that failed and should be further investigated. However, a requirement for this procedure is the availability of sufficient signals that correlate with the sensor to model.

An example is given in Figure 2.6. A software sensor to estimate $S_{\mathrm{NH4N,NB2/1}}$ was built with a GLSR model and Advanced Knowledge. The model included eight sensors and 79 variables (with transformations and lags). After 24 hours of validation, the oxygen sensor $C_{\mathrm{O2,NB2/2}}$ was set to measure a constant concentration of 0.5 g/m^3. Considering the estimates of the secondary sensors plotted in Figure 2.6, the sensor that failed can be identified. Only the model which was calibrated without using data of sensor $C_{\mathrm{O2,NB2/2}}$ neither exhibits a change in the width of the prediction interval, ΔPI, nor noticeable changes in the prediction after the failure was introduced.

Computations have shown that the same procedure can be applied if $C_{\mathrm{O2,NB2/2}}$ was subject to drift. However, the decrease of the accuracy and the increase of ΔPI is not abrupt as in the example above. As consequence, the detection of a drift can be slightly delayed.

2.4 Conclusions

This work investigated the applicability of data-driven modeling techniques to support WWTP operation. A simple procedure was introduced to systematically generate software sensors based on data available in the SCADA system of the plant. Practical deployment scenarios for software sensors include the creation of redundant measurements for sensor diagnosis and the replacement of existing sensors.

The procedure was tested with four different modeling techniques (GLSR, ANN, SOM and RF) with and without dimensionality reduction by PCA and by taking into account three different levels of expert knowledge in two full-scale experiments. Even with linear modeling techniques, it was possible to automatically generate accurate software sensors, despite the highly dynamic and non-linear hydraulic and biological WWTP processes.

Expert knowledge was important not only for the interpretation of generated software sensors but also for the reduction of the data set used for software-sensor generation to meaningful and reliable sensors. For the former, the models must be transparent and favorably parsimonious, which is satisfied only for SOM and GLSR models. without the application of PCA. For GLSR models, it was shown that software sensors have improved long-term accuracy with higher degrees of available expert knowledge and for longer calibration periods.

For safe deployment on-site, it is crucial that the sensors provide some measure of uncertainty for their predictions. By investigating the prediction intervals of the GLSR models or the quantification error of the SOMs, rare events and the failure of model input sensors for which the software-sensor prediction was unreliable could be identified.

Despite the fact that the data-driven models are not intelligent, with the right tools, they can reliably provide the operator with valuable information extracted from data already available in the SCADA system in a cost-effective way. In addition, the active exploration of the automatically generated models can increase understanding of in-plant processes.

2.5 Supporting information

Supporting content associated with this article can be found in Appendix B (page 99 ff.)

References

Aubrun, C., 2001. Software sensor design for COD estimation in an anaerobic fluidized bed reactor. Water Science and Technology 43 (7), 115–122.

Baeza, J. A., Gabriel, D., Lafuente, J., 2002. In-line fast OUR (oxygen uptake rate) measurements for monitoring and control of WWTP. Water Science and Technology 45, 19–28.

Belanche, L. A., Valdés, J. J., Comas, J., Roda, I. R., Poch, M., 1999. Towards a model of input-output behaviour of wastewater treatment plants using soft computing techniques. Environmental Modelling & Software 14 (5), 409–419.

Benazzi, F., Gernaey, K. V., Jeppsson, U., Katebi, R., 2007. On-line estimation and detection of abnormal substrate concentrations in WWTPS using a software sensor: A benchmark study. Environmental Technology 28 (8), 871–882.

Bernholt, T., Fried, R., Gather, U., Wegener, I., 2006. Modified repeated median filters. Statistics and Computing 16 (2), 177–192.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Carstensen, J., Harremoës, P., Strube, R., 1996. Software sensors based on the grey-box modelling approach. Water Science and Technology 33 (1), 117–126.

Carstensen, J., Madsen, H., Poulsen, N. K., Nielsen, M. K., 1994. Identification of wastewater treatment processes for nutrient removal on a full-scale WWTP by statistical methods. Water Research 28 (10), 2055–2066.

Cecil, D., Kozlowska, M., 2010. Software sensors are a real alternative to true sensors. Environmental Modelling & Software 25 (5), 622–625.

Céréghino, R., Park, Y.-S., 2009. Review of the Self-Organizing Map (SOM) approach in water resources: Commentary. Environmental Modelling & Software 24 (8), 945–947.

Corominas, L., Villez, K., Aguado, D., Rieger, L., Rosén, C., Vanrolleghem, P. A., 2011. Performance evaluation of fault detection methods for wastewater treatment processes. Biotechnology and Bioengineering 108 (2), 333–344.

Corona, F., Reinikainen, S.-P., Aaljoki, K., Perkiö, A., Liitiäinen, E., Baratti, R., Simula, O., Lendasse, A., 2008. Wavelength selection using the measure of topological relevance on the self-organizing map. Journal of Chemometrics 22 (11-12), 610–620.

Dellana, S. A., West, D., 2009. Predictive modeling for wastewater applications: Linear and nonlinear approaches. Environmental Modelling & Software 24 (1), 96–106.

Dewasme, L., Bogaerts, P., Vande Wouwer, A., 2009. Monitoring of bioprocesses: Mechanistic and data-driven approaches. Studies in Computational Intelligence 218, 57–97.

Gernaey, K. V., van Loosdrecht, M. C. M., Henze, M., Lind, M., Jørgensen, S. B., 2004. Activated sludge wastewater treatment plant modelling and simulation: state of the art: Environmental Sciences and Artificial Intelligence. Environmental Modelling & Software 19 (9), 763–783.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction, 2nd Edition. Springer series in statistics. Springer, New York, NY.

Henze, M., Gujer, W., Mino, T., van Loosdrecht, M., 2000. Activated sludge models ASM1, ASM2, ASM2d and ASM3. Scientific and technical report No. 9. IWA Publishing, London.

Hill, D. J., Minsker, B. S., 2010. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. Environmental Modelling & Software 25 (9), 1014–1022.

Jackson, J. E., 1991. A user's guide to principal components. John Wiley & Sons, New York.

James, S. C., 2000. On-line estimation in bioreactors: A review. Reviews in Chemical Engineering 16 (4), 311–340.

Jansson, A., Rottorp, J., Rahmberg, M., 2002. Development of a software sensor for phosphorus in municipal wastewater. Journal of Chemometrics 16 (8-10), 542–547.

Kalteh, A. M., Hiorth, P., Bemdtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling & Software 23 (7), 835–845.

Karim, M., Hodge, D., Simon, L., 2003. Data-Based Modeling and Analysis of Bioprocesses: Some Real Experiences. Biotechnology Progress 19 (5), 1591–1605.

Kim, M. H., Kim, Y. S., Prabu, A. A., Yoo, C. K., 2009. A systematic approach to data-driven modeling and soft sensing in a full-scale plant. Water Science and Technology 60 (2), 363–370.

Kohonen, T., 2001. Self-organizing maps, 3rd Edition. Springer, Berlin, London.

Langergraber, G., Fleischmann, N., Hofstadter, F., 2003. A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. Water Science and Technology 47 (2), 63–71.

Lee, D. S., Lee, M. W., Woo, S. H., Kim, Y.-J., Park, J. M., 2006. Nonlinear dynamic partial least squares modeling of a full-scale biological wastewater treatment plant. Process Biochemistry 41 (9), 2050–2057.

Lindberg, C.-F., Carlsson, B., 1996. Estimation of the respiration rate and oxygen transfer function utilizing a slow DO sensor. Water Science and Technology 33 (1), 325–333.

Luccarini, L., Bragadin, G. L., Colombini, G., Mancini, M., Mello, P., Montali, M., Sottara, D., 2010. Formal verification of wastewater treatment processes using events detected from continuous signals by means of artificial neural networks. Case study: SBR plant. Environmental Modelling & Software 25 (5), 648–660.

Masson, M. H., Canu, S., Grandvalet, Y., Lynggaard-Jensen, A., 1999. Software sensor design based on empirical data. Ecological Modelling 120 (2-3), 131–139.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2006. Introduction to linear regression analysis, 4th Edition. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, NJ.

Mouton, A. M., Alcaraz-Hernández, J. D., Baets, B. d., Goethals, P. L. M., Martínez-Capel, F., 2011. Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. Environmental Modelling & Software 26 (5), 615–622.

Paliwal, M., Kumar, U. A., 2011. Assessing the contribution of variables in feed forward neural network. Applied Soft Computing 11 (4), 3690–3696.

Park, Y.-S., Tison, J., Lek, S., Giraudel, J.-L., Coste, M., Delmas, F., 2006. Application of a self-organizing map to select representative species in multivariate analysis: A case study determining diatom distribution patterns across France. Ecological Informatics 1 (3), 247–257.

Raduly, B., Gernaey, K. V., Capodaglio, A. G., Mikkelsen, P. S., Henze, M., 2007. Artificial neural networks for rapid WWTP performance evaluation: Methodology and case study. Environmental Modelling & Software 22 (8), 1208–1216.

Rieger, L., Siegrist, H., Winkler, S., Saracevic, E., Votava, R., Nadler, J., 2002. In-situ measurement of ammonium and nitrate in the activated sludge process. Water Science and Technology 45 (4-5), 93–100.

Rieger, L., Thomann, M., Joss, A., Gujer, W., Siegrist, H., 2004. Computer-aided monitoring and operation of continuous measuring devices. Water Science and Technology 50 (11), 31–39.

Rosen, C., Lennox, J. A., 2001. Multivariate and multiscale monitoring of wastewater treatment operation. Water Research 35 (14), 3402–3410.

Rosen, C., Röttorp, J., Jeppsson, U., 2003. Multivariate on-line monitoring: challenges and solutions for modern wastewater treatment operation. Water Science and Technology 47 (2), 171–179.

Solomatine, D., See, L. M., Abrahart, R. J., 2008. Data-Driven Modelling: Concepts, Approaches and Experiences. In: Abrahart, R. J., See, L. M., Solomatine, D. P. (Eds.), Practical Hydroinformatics. Vol. 68. Springer, Berlin, Heidelberg, pp. 17–30.

Sotomayor, O. A., Park, S. W., Garcia, C., 2002. Software sensor for on-line estimation of the microbial activity in activated sludge systems. ISA Transactions 41 (2), 127–143.

Verikas, A., Gelzinis, A., Bacauskiene, M., 2011. Mining data with random forests: A survey and results of new tests. Pattern Recognition 44 (2), 330–349.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11 (3), 586–600.

Yoo, C. K., Lee, J.-M., Lee, I.-B., Vanrolleghem, P. A., 2004. Dynamic monitoring system for full-scale wastewater treatment plants. Water Science and Technology 50 (11), 163–171.

Chapter 3

Identification of industrial wastewater by clustering wastewater treatment plant influent ultraviolet visible spectra

Water Science and Technology, 2011, 63(6), 1153-1159

David J. Dürrenmatt and Willi Gujer

Identification of industrial wastewater by clustering wastewater treatment plant influent ultraviolet visible spectra

David J. Dürrenmatt and Willi Gujer**

* *Institute of Environmental Engineering, ETH Zurich, 8093 Zurich, Switzerland and Swiss Federal Institute of Aquatic Science and Technology, Eawag, 8600 Dübendorf, Switzerland (E-mail: david.duerrenmatt@eawag.ch)*

Abstract A procedure is proposed which allows the detection of industrial discharge events at the inlet of a wastewater treatment plant without the need for measurements performed at the industry, for special equipment and for exact knowledge of the industrial sewage. By performing UV/Vis measurements at the inlet of a plant and analyzing them with a two-staged clustering method consisting of the self-organizing map algorithm and the Ward clustering method, typical sewage clusters can be found. In an experiment performed at a mid-sized Swiss plant, one cluster of a cluster model with five clusters could be attributed to an industrial laundry. Out of 95 laundry discharging events measured in a validation period, 93 were correctly detected by the proposed algorithm, two were false positives and five were false negatives.

Keywords data mining; industrial wastewater; self-organizing maps; UV/Vis photospectrometry; Ward clustering

3.1 Introduction

Industrial wastewater can have a significant impact on the performance of a wastewater treatment plant (WWTP). Discharged at times when the plant is running at full capacity, e.g. during peak hours, industrial sewage can cause overloading and thus exceeding effluent concentration constraints.

WWTP operators are generally not provided relevant information by the industrial sites in the catchment area, neither on the type of sewage, the temporal pattern of discharging, nor on the exact load which is released to the sewer system, although this information is important for optimal plant control. Analyzing WWTP inflow with focus on the detection of industrial dischargers and the attribution of a discharge event to its producer, on the other hand, is difficult.

Methods for the detection of unusual changes in the wastewater composition and abnormal influent characteristics can be found in the literature. Langergraber *et al.* (2004), for example,

present a method to generate alarm parameters from measured influent UV/Vis absorption spectra. Although their method can be used for early warning, it does not directly provide information on the discharger. A method to search the sources of wastewater which inhibits nitrification is given by Kroiss *et al.* (1992); the method however requires sampling of the industrial sewages. This paper, in contrast, proposes a two-staged clustering approach which uses influent UV/Vis absorption spectra only to reveal information on the wastewater producers in a catchment area and on the pattern of their discharging, which is novel. Clustering is a fundamental task in data mining and aims at grouping data instances into subsets such that similar instances are grouped together (Rokach and Maimon, 2005). Using a different approach to cluster UV/Vis spectra measured at a fuel park WWTP, Lourenço *et al.* (2006) show that information can be extracted from the spectra and can be used for qualitative monitoring. However, their sample size was small and the spectra were sampled at several locations within the plant, thus exhibiting clear differences.

The proposed approach consists of a self-organizing map (SOM) which is used to generate a smaller but still representative data set using preprocessed UV/Vis absorption spectra. In the second step, the reduced data set is clustered by Ward's hierarchical agglomerative clustering algorithm and the clusters are manually labeled. The installation of an UV/Vis sensor at the WWTP inlet is sufficient (and may serve other purposes) and there is no need for the industry to install special measuring equipment, data storage and transmission infrastructure. The deployed clustering model detects and distinguishes different sewages and attributes them to producers.

3.2 Material and methods

3.2.1 In-situ UV/Vis photospectrometry

UV/Vis photospectrometry measures the absorption of light from the ultraviolet to the visible range. Although there are methods which allow the extraction of quantitative information on concentrations of chemical compounds which absorb in the given wavelength range, they cannot directly be applied for wastewater analysis because of physical or chemical interference (Thomas and Cerda, 2007). However, taking into account that a UV/Vis absorption spectrum is unique for a certain sewage composition, it can be considered as a fingerprint and be used for further analysis.

The spectra for this study were recorded with a submersible spectrometer probe (spectro::lyzer by s::can Messtechnik GmbH, Vienna, Austria) with 5 mm optical path length which measures the turbidity compensated absorbance between the wavelengths of 200 and 742.5 nm in 2.5 nm steps (for more information on the sensor, see e.g. Langergraber *et al.*, 2003).

3.2.2 Site description

This experiment is performed in a Swiss community with about 20'000 population equivalents. The average discharge originating from the catchment at the WWTP inlet is 80 L/s. An industrial laundry is situated in the catchment approximately 1 km upstream of the plant, which corresponds to an estimated sewer flow time of 26 minutes. During the irregularly

Table 3.1: Characterization of domestic and laundry wastewater (sampled on 25thMay 2010 12:00).

Parameter	Laundry	Domestic
CODeq [mg/L]	1970	800
Soluble CODeq [mg/L]	1556	290
NH4-N [mg/L]	1.1	25
NO3-N [mg/L]	2.8	0.3
NO2-N [mg/L]	0.45	0.04
PO4-P [mg/L]	0.5	2.8
pH [-]	8.1	7.2
Conductivity [mS/cm]	1.5	1.3

occurring discharge events, 10 L/s of laundry sewage are pumped into the sewer system for a variable amount of time (18 minutes on average). The composition of laundry wastewater is compared to domestic wastewater in Table 3.1.

In the process information system of the plant, data of common operating parameters are readily available. In a measuring campaign from 1st June 2009 00:00 to 30th June 2009 24:00, a total of 31,614 UV/Vis absorption spectra were recorded after the fine screen (6 mm) with a sampling interval of 1 minute.

At the laundry, the discharge events were recorded in a measuring campaign from 2nd June 2009 00:00 to 12th June 2009 12:00. To be used for model validation, the events measured at the laundry must be synchronized with the events detected at the WWTP by taking into account the flow in the sewer. This is done by adding the estimated mean sewer flow time to the recorded events.

3.2.3 Two-stage clustering approach: self-organizing maps and Ward clustering

UV/Vis data has high dimensionality, is subject to significant noise and suffers from outliers, this makes clustering difficult. Thus we have chosen a two-staged approach which has proven powerful especially when dealing with "noisy and messy" data (Canetta *et al.*, 2005; Vesanto and Alhoniemi, 2000). The approach is illustrated in Figure 3.1.

Data preparation

Because the concentrations of the domestic sewage and the laundry wastewater vary over time, a preprocessing method has been developed which reduces dilution phenomena of a UV/Vis fingerprint while emphasizing its characteristic shape. This is achieved by normalizing the absorption spectra $a(\lambda)$ (absorbance a at wavelength λ) to have zero mean and unit variance and shifting it so that all measured absorption rates of wavelengths greater than a wavelength λ_c are aligned. Mathematically speaking, spectra are transformed by the non-intuitive equation

$$\tilde{a}(\lambda) = \frac{1}{\sqrt{\mathrm{var}(a)}} \left(a(\lambda) - \frac{\int_{\lambda_c}^{\infty} a(\lambda)\mathrm{d}\lambda}{\int_{\lambda_c}^{\infty} \mathrm{d}\lambda} \right). \tag{3.1}$$

After transformation, all $\tilde{a}(\lambda)$ with $\lambda \geq \lambda_c$ are cropped.

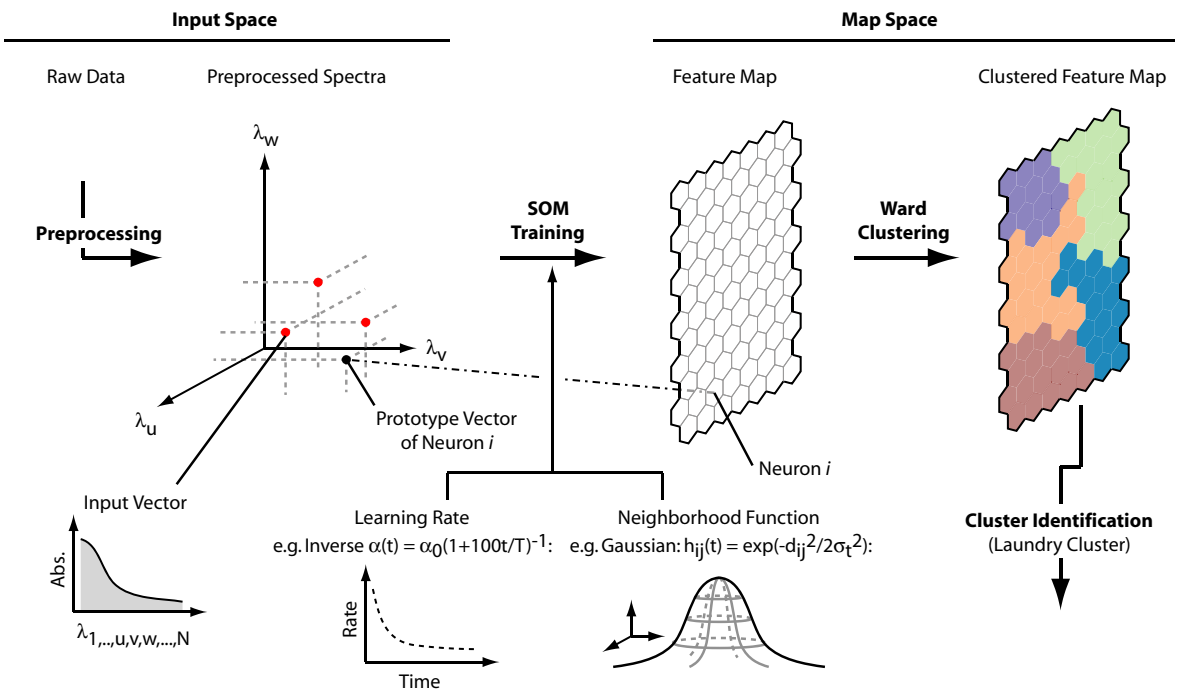

Figure 3.1: Scheme illustrating the two-staged clustering approach: Preprocessed UV/Vis spectra (one spectrum corresponds to one input vector) which form a high-dimensional input space are mapped to a two dimensional SOM feature map in the map space. To train the SOM, i.e. to find the optimum prototype vectors for all neurons, a learning rate function (e.g. inverse where α_0 denotes the initial learning rate and T the training length) and a neighborhood function (e.g. Gaussian where d_{ij} denotes the distance between neurons i and j on the feature map and σ_t the neighborhood radius) are applied. The trained feature map is clustered by the Ward Clustering algorithm and cluster labels are manually assigned.

The preprocessed spectra are split into a calibration set, which contains the UV/Vis data from 13th to 30th June 2009 (no laundry data available) and a validation set, which contains the UV/Vis data and laundry discharge data from 2nd to 12th June 2009: The former is used to build the model and to detect and label the clusters (see below), the latter to assess the model performance.

Self-organizing maps

The self-organizing maps are a variant of artificial neural networks based on unsupervised learning, originally proposed by Kohonen (2001). They learn to cluster groups of similar input data in a non-linear projection from a high-dimensional data-space onto a lower-dimensional discrete lattice of neurons on an output layer (feature map, cf. Figure 3.1) in an orderly fashion (Kalteh *et al.*, 2008). This is done in a topology preserving way which means that neurons physically located close the each other have similar input patterns. Additionally, the SOM is tolerant to noise which is especially helpful when dealing with experimental measurements.

Each neuron has assigned a prototype vector having the same dimensionality as the input data. During training, these vectors are optimized in order to represent the whole input data; the set of prototype vectors is therefore representative for the data set. The optimization of the prototype vectors is proportional to a learning rate and a neighborhood function, both monotonically decreasing during the ordering process. The former is a scalar; the latter forms a smoothing kernel around the prototype vector and makes sure that only input vectors within a certain neighborhood affect the prototype vector. Both are important to allow a regular smoothing.

The prototype vector w_i of neuron i which has minimum distance to an input vector $\tilde{a}_t$ is the winning neuron for this input vector and is called the best-matching unit (BMU). The distance between the input vector and its BMU is called quantification error and given by the following Euclidian distance:

$$q.e. = \|\tilde{a}_t - w_i\| \tag{3.2}$$

In this study, the software SOM toolbox for Matlab (Vesanto *et al.*, 2000) was used to train and evaluate the SOMs.

Ward clustering

When applying the hierarchical agglomerative Ward clustering method (Ward, 1963) on the SOM prototype vectors, each vector first forms its own cluster. Then, subsequently, the two clusters with minimum Ward distance are merged (the Ward distance between two clusters is defined as amount of variance added when two clusters are merged). The aim is to have small variance within a cluster and high variance between the clusters.

The optimum number of clusters is estimated by taking into account the Davies-Bouldin index (db-index; Davies and Bouldin, 1979), which is the averaged similarity between each cluster and its most similar one (Halkidi *et al.*, 2002). Because clusters with minimum similarity are aimed, the db-index is minimized.

The task of the expert is it now to identify the cluster in which the UV/Vis spectra of laundry/domestic wastewater mixtures lie (hereafter named qlaundry cluster).

3.3 Results and discussion

3.3.1 Clustering model

The clustering model given in Figure 3.2 was trained following the approach illustrated in Figure 3.1 using the preprocessed spectra of the calibration set (17,215 measurements, 54% of the data set, cropped at $\lambda_c = 324$nm). A SOM with a two dimensional feature map (the neurons are arranged in a hexagonal grid) with map size 50x13 was trained using a Gaussian neighborhood function with σ_t linearly decreasing with time from 7 to 1 and an inverse learning rate function with $\alpha_0 = 0.5$ (see Figure 3.1). Please note that for all distance measures, the Euclidian distance was used.

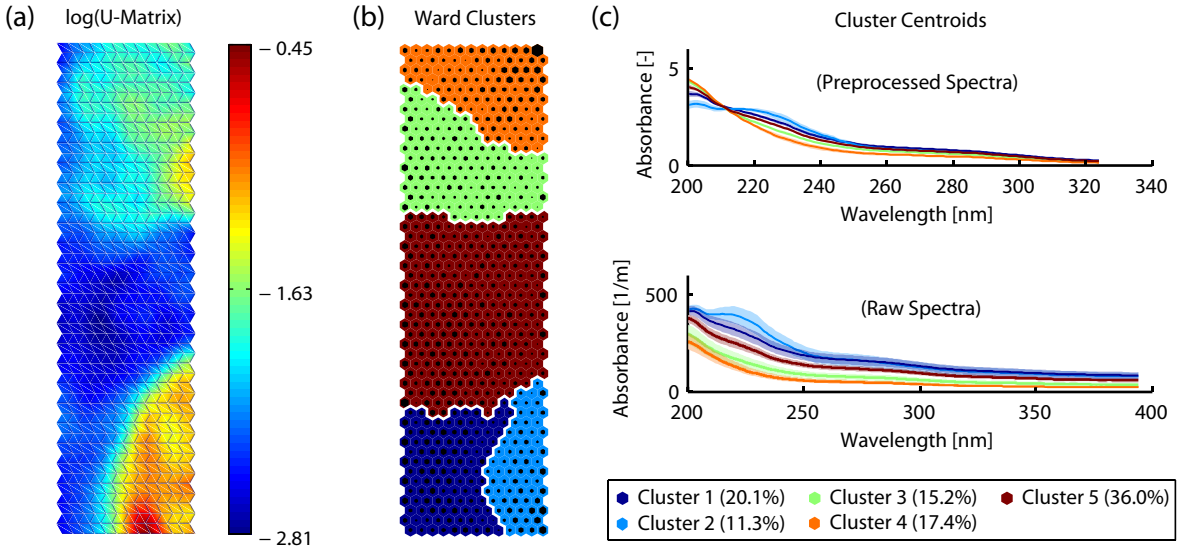

Figure 3.2: Feature maps of the trained SOM are presented in (a) and (b). The colorization of the feature map in (a) indicates the average distance of a neuron to its neighbors (the logarithm of the so-called U-Matrix). In (b), the colored areas represent the clusters found by Ward clustering and the size of the black hexagons within the neurons the count of input vectors for which the neuron is the BMU. In (c), the Ward cluster centroids are plotted; the percentages of spectra in each cluster are given in the legend.

The topographic error (the proportion of all data vectors whose first and second BMUs are not adjacent vectors, cf. Kohonen, 2001) of the SOM is 0.064 and the average quantization error 0.153. The Ward algorithm was applied for a cluster size of five, which had the lowest db-index of 0.67.

3.3.2 Laundry cluster detection

In order to find the cluster which contains BMUs for the UV/Vis spectra measured when laundry wastewater is being discharged, the use of three visualizations was advantageous:

i) The plot of the cluster centroids (cf. Figure 3.2c) is helpful to detect clusters significantly deviating from the others and having a shape consistent with available prior knowledge. Cluster 2 deviates from the others.

ii) Considering a time series plot showing the particular cluster overlaid with the measured inflow parameters (Figure 3.3a), one would again select Cluster 2 (higher temperature, slightly elevated discharge).

iii) A ring map revealing the periodicity of the discharging (Figure 3.3b) exhibits an obvious, albeit irregular, operational schedule of Monday-Friday 7:00 to 23:00 and frequently of Saturday 7:00 to 12:00 for Cluster 2, which corresponds to the production schedule of the laundry.

As a result, one can say that laundry discharging events are contained in Cluster 2 with high probability. It is important to notice that the quantification error plotted in Figure

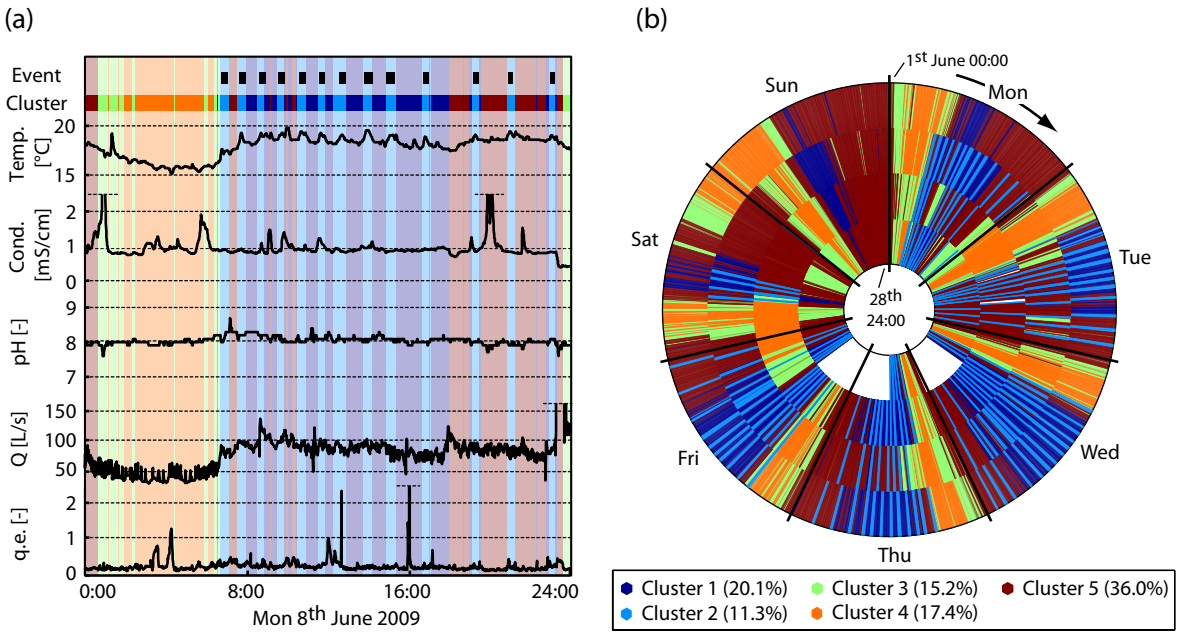

Figure 3.3: (a) Discharging events measured at the industrial site shifted by the mean residence time in the sewer (filled rectangles), clusters assigned to the UV/Vis measurements, operating data measured at the inlet and the quantification error (q.e.) of the SOM for the period of one day. (b) Ring map indicating the weekly periodicity of the clusters (1^{st}June: Holiday).

3.3a remains small during most of the detected events which shows that the UV/Vis spectra measured during an event are appropriately represented on the SOM.

To set up the model and select the associated cluster no quantitative information on the time and duration of the discharging events is needed. In the next section, the clustering model will be validated using quantitative information recorded at the industrial site.

Given the case that the cluster cannot be identified, heavy dilution effects, insufficient absorption of characteristic compounds or interference due to mixtures of several similar sewages could be hindering reasons. In bigger catchment areas where there are many industries with interfering spectra the integration of other operating data or mounting the measuring devices upstream of the plant might help.

3.3.3 Model performance

The model performance is assessed by evaluating the clustering model with the validation data set (14,399 spectra, 46% of the data set) and comparing the time periods when Cluster 2 is detected with the measured discharging events at the laundry shifted by the mean flow time in the sewer.

The clustering model predicts a total of 95 events. Comparing these to the 98 events measured, 93 were assigned correctly, two were false positives and five were false negatives (cf. Figure 3.4a). This corresponds to a failure rate of 7%.

The average duration is 18.0 minutes for the measured and 20.8 minutes for the predicted events. The error of the predicted event duration is thus 2.8 ± 3.9 minutes (mean ± standard

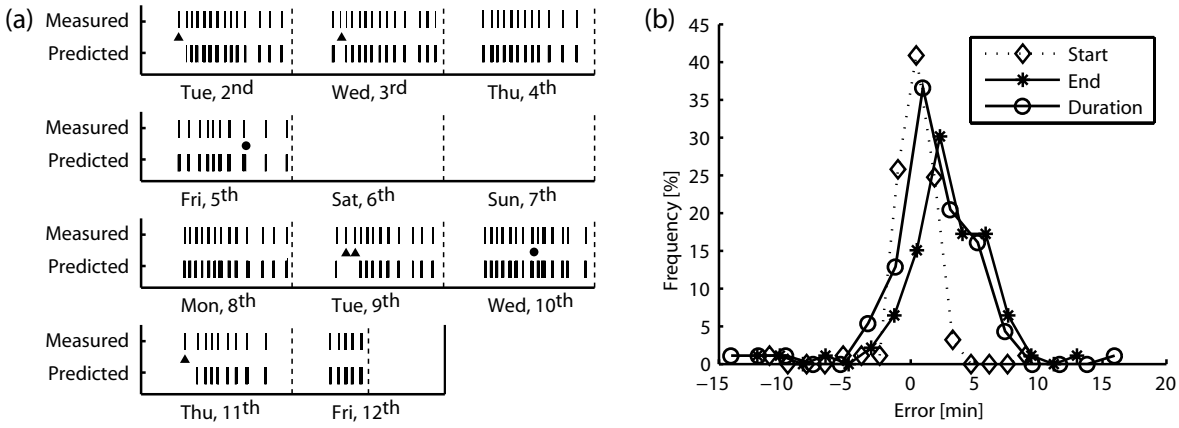

Figure 3.4: (a) Measured and predicted events for the validation period (the width of the bar corresponds to the duration of the event; bold circles indicate false positives, triangles false negatives). (b) Distribution of the error when comparing the predicted with the measured events.

deviation); the error of the predicted event start is 0.9 ± 2.0 minutes and 3.7 ± 3.6 minutes of the event end (the distribution is given in Figure 3.4b). Some of this error may be caused by neglecting the variable flow velocity and dispersion in the sewer when shifting by the mean sewer flow time.

3.3.4 Detection limit

The detection limit of the clustering model for the detection of laundry discharge events depends on the dilution of laundry and domestic sewage at the WWTP inlet.

The maximum dilution of laundry sewage which can still be detected by the clustering model was experimentally estimated by measuring and clustering the UV/Vis spectra of different dilutions of domestic and laundry sewage (both grab-sampled on 25thMay 2010 12:00). Dilutions which contained more than 7% laundry sewage were detectable. This corresponds to a minimum fraction of the COD load of 12%. Comparing the discharge of 10 L/s of the laundry with the discharge at the inlet of the WWTP which is 124 L/s (85% percentile), one can conclude that the majority of the discharging events are detectable.

It was observed that dilution caused by rainfall events only has minor effects on the model performance. This can be explained: Although pollutant concentrations are lower during rainfall events, the ratio between domestic and laundry sewage approximately remains the same. Five laundry discharging events out of six which occurred during three rainfall events (WWTP inflow greater than 200 L/s) in the validation period were detected correctly.

3.3.5 Long term validity

To ensure long term validity, it must be ascertained that the conditions under which the clustering model was calibrated remain constant. That is, the quality of the laundry sewage does not change, e.g. due to changed processes, and that there is no other interfering sewage

originating from the catchment area which exhibits a similar UV/Vis fingerprint and activates a BMU in the same cluster on the feature map of the SOM.

To detect possible changes, it is advisable to *i)* track the quantification error during clustering (high quantification errors indicate novel / abnormal patterns) and to *ii)* recompute the cluster model regularly and check whether the number of clusters remains the same and the clustering results are similar (disappearing and appearing clusters indicate changes in the catchment area). Recomputing can be highly automated and does not need additional measurements to be performed. However, when changes are detected and validation is advisable, additional measurements are required.

3.4 Potential applications

The authors see four practical tasks for which the proposed method is helpful: *i)* The localization of an unknown sewage producer (by inferring rules from the clustered time series and by estimating sewer flow distance by analyzing the trajectories on the feature map), *ii)* the verification of legal compliance (i.e. if a producer violates agreed discharging loads), *iii)* the implementation of source related cost-allocation relying on detected events and *iv)* the use of the SOM to detect abnormal sewage compositions (by analyzing the quantification error).

3.5 Conclusions

Using a robust and reliable UV/Vis sensor mounted at the inlet of a WWTP in combination with the proposed two-staged clustering approach (SOM in combination with Ward's clustering algorithm), it is possible to detect and distinguish different sewage compositions, thus reveal information about the catchment area. If the sewage types can further be linked to their producers the temporal pattern of discharging can be visualized and used for further investigations. In the given example, the approach generated a SOM with five characteristic clusters, of which one could be assigned to an industrial laundry. Model validation revealed that out of 95 events of variable duration which occurred in 12 days, 93 could be detected with two false positives and five false negatives only.

It has to be stressed that some of the simplifications and assumptions which were justified in this example do not hold in more complex catchments. For instance, having two industrial sites A and B in the catchment which produce sewages whose UV/Vis spectra do not interfere, thus whose discharging can be differentiated by the SOM, one already has to identify four clusters: "A", "B", "A and B" and "none".

3.6 Acknowledgements

The WWTP staff is acknowledged for providing open access to their SCADA system and Michael Burckhardt for the provision of the laundry discharge data.

References

Canetta, L., Cheikhrouhou, N., Glardon, R., 2005. Applying two-stage SOM-based clustering approaches to industrial data analysis. Production Planning & Control 16 (8), 774–784.

Davies, D. L., Bouldin, D. W., 1979. A cluster separation measure. Ieee Transactions on Pattern Analysis and Machine Intelligence 1 (2), 224–227.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2002. Clustering validity checking methods: Part II. Sigmod Record 31 (3), 19–27.

Kalteh, A. M., Hiorth, P., Bemdtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling & Software 23 (7), 835–845.

Kohonen, T., 2001. Self-organizing maps, 3rd Edition. Springer, Berlin, London.

Kroiss, H., Schweighofer, P., Frey, W., Matsche, N., 1992. Nitrification inhibition - a source identification method for combined municipal and or industrial waste-water treatment plants. Water Science and Technology 26 (5-6), 1135–1146.

Langergraber, G., Fleischmann, N., Hofstadter, F., 2003. A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. Water Science and Technology 47 (2), 63–71.

Langergraber, G., Weingartner, A., Fleischmann, N., 2004. Time-resolved delta spectrometry: a method to define alarm parameters from spectral data. Water Science and Technology 50 (11), 13–20.

Lourenço, N. D., Chaves, C. L., Novais, J. M., Menezes, J. C., Pinheiro, H. M., Diniz, D., 2006. UV spectra analysis for water quality monitoring in a fuel park wastewater treatment plant. Chemosphere 65 (5), 786–791.

Rokach, L., Maimon, O. Z., 2005. Clustering Methods. In: Maimon, O. Z., Rokach, L. (Eds.), Data mining and knowledge discovery handbook. Springer, Ramat-Aviv.

Thomas, O., Cerda, V., 2007. From spectra to qualitative and quantitative results. In: O. Thomas and C. Burgess (Ed.), UV-Visible Spectrophotometry of Water and Wastewater. Vol. 27. Elsevier, pp. 21–45.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11 (3), 586–600.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 2000. SOM Toolbox for Matlab 5. Espoo, Finland.

Ward, J. H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58 (301), 236–244.

Chapter 4

New method for the assessment of the discharge distribution at hydraulic flow dividers

In Preparation

David J. Dürrenmatt, Samuel Zeller and Willi Gujer

New method for the assessment of the discharge distribution at hydraulic flow dividers

David J. Dürrenmatt[a,b,*], *Samuel Zeller*[a] *and Willi Gujer*[b]

[a] *Institute of Environmental Engineering, ETH Zurich, 8093 Zurich, Switzerland*
[b] *Swiss Federal Institute of Aquatic Science and Technology, Eawag, 8600 Dübendorf, Switzerland*
* *Corresponding author phone: +41 44 823 5407; fax: +41 44 823 5389; email: david.duerrenmatt@eawag.ch*

Abstract This paper introduces a simple and practical method that can quantify the discharge distribution at flow dividers commonly implemented in wastewater treatment plants (WWTPs) to evenly feed reactor units of treatment steps operated in parallel. Although the flow dividers are designed to split the flow evenly, they often fail to do so. Uneven distribution can have severe consequences on plant performance but also poses a problem for WWTP modeling when even splitting has to be assumed due to the lack of detailed information on the discharge distribution. The presented method compares any signal measured at the flow divider that shows variability to other signals of the same type measured in the effluents of the reactor units operated directly downstream of the divider. By assigning patterns found in the influent to the corresponding patterns in the effluents using a dynamic time warping algorithm, the discharge distribution can be determined. A theoretical analysis reveals that the accuracy of the method increases with decreasing dispersion and with decreasing reactions that affect the measured signal in the reactors downstream of the flow divider; it also shows that the method has a wide area of application for WWTPs. For both synthetic and real experiments, it is shown that the relative systematic error in the discharge signal under the assumption of equal distribution could be reduced by more than 50% in average when correcting the assumption of equal discharge distribution.

Keywords discharge distribution; flow divider; plant performance; reactor hydraulics; modeling; dynamic time warping

Nomenclature

C_0	Signal measured in the influent ($M\ L^{-3}$, $^\circ C$)
C_j	Signal measured in reactor j, $0 < j \leq N$ ($M\ L^{-3}$, $^\circ C$)
$\mathbf{D}$	Distance matrix
$E_{rel,\theta}$	Theoretical relative error of θ
$E_{rel,\xi}$	Theoretical relative error of ξ
f	Frequency (T^{-1})
f_{cut}	Cut-off frequency of low-pass filter (T^{-1})

k	Reaction constant (T^{-1})
L	Data length (T)
N	Number of reactors in a tanks-in-series model (dimensionless)
P	Number of parallel lanes (dimensionless)
Q_a	Flowrate assuming uniform distribution ($L^3\ T^{-1}$)
Q_i	Flowrate to lane i ($L^3\ T^{-1}$)
Q_{tot}	Total influent discharge ($L^3\ T^{-1}$)
R	Number of runs (dimensionless)
r	Reaction rate ($M\ L^{-3}\ T^{-1}$, $°C\ T^{-1}$)
T_0	Influent temperature (°C)
T_i	Effluent temperature of lane i (°C)
V	Reactor volume (one lane) (L^3)
W	Warping path (dimensionless)
$\bar{W}$	Averaged warping path (dimensionless)
Z	Order of polynomial (dimensionless)
Δt	Time step (T)
σ_m	Standard deviation of measurement error
θ	Hydraulic residence time (T)
θ_{mean}	Mean hydraulic residence time (T)
$\hat{\theta}$	Estimated hydraulic residence time (T)
$\theta_{max,est}$	Estimated maximum hydraulic residence time (T)
ξ_i	Relative deviation of Q_i from Q_a
$\hat{\xi}_i$	Relative deviation of Q_i from Q_a as polynomial function

4.1 Introduction

In many wastewater treatment plants (WWTPs), several treatment steps are implemented by multiple identical reactor units operated in parallel; this system offers optimal efficiency and the ability to disconnect individual reactors for maintenance. The flow is split by hydraulic flow dividers that are intended to provide equalized charging of the subsequent units.

However, these devices are often inaccurate (Patel *et al.*, 2008), which leads to uneven loading that can result in performance losses, especially at peak loads. The causes of unequal discharge distribution include the design of the flow divider, installation that is not exactly level and bends upstream of the device (Dutta *et al.*, 2010; Patel *et al.*, 2008). Even though the effects are often visible to the human eye, their quantitative assessment is not trivial; only in special cases are there flow measurements at each of the branches.

The operational problems associated with asymmetric distribution include the following: *i)* effluent constraint violations, which occur when activated sludge tanks are unevenly fed and the flow to one or more exceeds the maximum specifications; and *ii)* inefficient plant control, which can occur when a sensor is mounted in only one of the several reactors that are operated in parallel.

In addition to the operational issues, the effects of asymmetric splitting can also be of significance for scientists and engineers. When performing measurements on WWTPs for modeling, it is often assumed that the discharge is evenly distributed among the different lanes operated in parallel. This assumption is necessary due to the lack of methods for a practical quantitative assessment of the discharge distribution. Considering that commonly used flow measurement techniques are already erroneous (see Harremoes *et al.*, 1993), another significant error is introduced when uniform distribution is assumed. This error poses a problem for successful modeling because an appropriate description of the hydraulic processes is crucial, as shown in Orhon *et al.* (1989).

Thus, coping with asymmetric distributions requires a feasible and cost-efficient method of determining the quantitative relationship between the discharge distribution and the total inflow.

4.1.1 Commonly used methods

Although there are established methods for discharge measurement and for investigation of the hydraulic behavior of a reactor that can be used for the determination of discharge distribution, their applicability in practice is often hindered.

A common method for the ad-hoc assessment of inequality in splitting is the mobile discharge measurement (see, for example, Hager, 1999). However, the accuracy of a chosen measurement method depends strongly on the constructional constraints and hydraulic conditions (Quevauviller *et al.*, 2006). The systematic measurement error of the flow rate in 18 WWTPs was evaluated by Port (1994), and an average error of -8.6% was found.

The conventional tracer experiment is a popular tool for providing insights into aspects of reactor hydraulics such as the residence time distribution (Gujer, 2008). Given the mean hydraulic residence time (HRT), θ_{mean}, and the volume of the reactor, V, the mean discharge can be calculated by

$$Q_{mean} = \frac{V}{\theta_{mean}} \tag{4.1}$$

In practice, tracer experiments are costly and labor-intensive (Ahnert *et al.*, 2010). Further difficulties arise by high flow variability, required mixing lengths and density flow. Even under optimal conditions, a residence time distribution is only valid for a specific discharge. Reactors that exhibit different behaviors for low and high flows require performing tracer experiments for each specific discharge.

More recent publications suggest the use of (naturally occurring) reactive tracers (Ahnert *et al.*, 2010; Gresch *et al.*, 2010). In Ahnert *et al.* (2010), the authors show that the use of temperature as a natural tracer, combined with a temperature model, allows the estimation of HRT distributions and provides a viable alternative to conventional tracer experiments. Keeping the flowrate constant by using storage basins and focusing on temperature peaks from cold stormwater events has been suggested. However, this method is not suitable for the quantification of discharge distributions as a function of the total discharge.

A third approach to determining the extent of the inequalities lies in the use of computational fluid dynamics (CFD) simulations to analyze the flow divider (Dutta *et al.*, 2010). Because

CFD requires a detailed geometrical representation and because already minor and possibly unknown effects can lead to asymmetries, the applicability of this method is limited.

In practice, what often remains is the assumption of an equal distribution.

4.1.2 Objective of this paper

This paper presents a novel procedure for estimating the discharge distribution of hydraulic flow dividers by comparing the influent and effluent measurements of the reactors directly downstream of a device using dynamic time warping (DTW), a dynamic programming method. The procedure requires easily measurable data that can be gathered with generally available low-maintenance sensors (e.g., temperature probes). The information on the discharge distribution that can be obtained using this method is valuable for the scientist, the engineer and the plant operator in the following situations:

- The flow divider is seen to be asymmetric. There is a gate valve at each branch that may be manually adjusted. By first assessing the discharge distribution and then iteratively adjusting the valve and re-assessing the situation, the valve position leading to the most uniform discharge distribution can be found.
- The plant is running at full capacity. Assessing the discharge distribution of a flow divider reveals that a more uniform distribution would more evenly utilize the available resources.
- A treatment step needs to be modeled for optimization studies. The modeler is aware of the unequal distribution among the parallel lanes. He or she applies the discharge distribution to get an estimate for the reactor inflow that is more accurate than the equal distribution assumption.

This paper is organized in the following manner. First, the physical background of the method is presented, and the area of application is specified by considering a theoretical system that has a closed-form solution. Next, a detailed procedure for performing a discharge distribution analysis in practice is presented. Finally, the results of applying the procedure to a wide range of synthetic systems and a real mid-sized Swiss WWTP are presented and discussed.

4.2 Method

Assume a treatment step operated in P parallel lanes and a flow divider that splits the total influent discharge Q_{tot} into P parts. The i-th lane is fed a discharge of

$$Q_i(t) = \frac{1+\xi_i}{P} Q_{tot}(t) \tag{4.2}$$

as a function of time t, where ξ_i is a factor quantifying the relative deviation from the uniform distribution that is defined by

$$\xi_i = \frac{PQ_i - Q_{tot}}{Q_{tot}}. \tag{4.3}$$

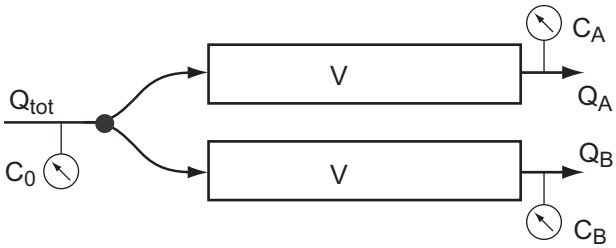

Figure 4.1: Hydraulic distribution device (black circle) followed by two reactors, each having a volume V. The total influent discharge is split into two parts, Q_A and Q_B. Three measuring locations are indicated: C_0 (influent), C_A (effluent A) and C_B (effluent B).

The mass balance equation, $Q_{tot} = \sum_{i=1}^{P} Q_i$ (and consequently $\sum \xi_i = 0$), also applies. Let Q_a be the discharge to every branch when a uniform distribution is assumed (i.e., when all the branches of the distribution box receive the same discharge). Thus, with $\xi_i = 0 \; \forall \; i$,

$$Q_a(t) = \frac{1}{P} Q_{tot}(t). \tag{4.4}$$

In principle, the objective of this paper is to find ξ_i, which can be expressed as a function of t or of Q_{tot}.

For simplicity, we will only consider treatment steps operated in two identical parallel lanes ($P = 2$), as illustrated in Figure 4.1. However, all the equations can be generalized for the case of $P \geq 2$. For the two branching flows, Q_A and Q_B, Eq. (4.3) can be written as

$$\xi_A = -\xi_B = \frac{Q_A - Q_B}{Q_A + Q_B} = \frac{\theta_B - \theta_A}{\theta_B + \theta_A} \tag{4.5}$$

by taking into account Eq. (4.1) and assuming that the reactor volumes are identical. θ_A and θ_B are the current mean hydraulic residence times in the respective reactor.

To estimate θ_A and θ_B, a method that has similarities to a tracer experiment is applied. However, instead of requiring injection of a known mass of a tracer substance at the influent of an observed system and measurement of its concentration at the effluent of each of the lanes, this method tracks the characteristic patterns that naturally occur in an influent signal $C_0(t)$ and assigns these patterns to the associated patterns in the two effluent series $C_A(t)$ and $C_B(t)$. The method further assumes that there is a relationship between θ_A and θ_B with the travel time between the observation of a pattern in the influent and in the effluent.

This principle is illustrated in Figure 4.2 by means of an ideal plug-flow reactor. "Water packet" I is observed in the influent at time t_0 and two time steps Δt later in the effluent. Hence, the observed travel time is $\hat{\theta}_A = 2\Delta t$. Similarly, the travel time is $2\Delta t$ for packet II, $3\Delta t$ for packet III and $4\Delta t$ for packet IV. The discharge through the reactor varies with time and is not known. Let $Q = (Q_0, Q_1, \ldots, Q_k, \ldots, Q_L)$ be the discharge at time $t = (t_0, t_1, \ldots, t_k, \ldots, t_L)$. For each observed travel time $\hat{\theta}$, a mass balance can be formulated as

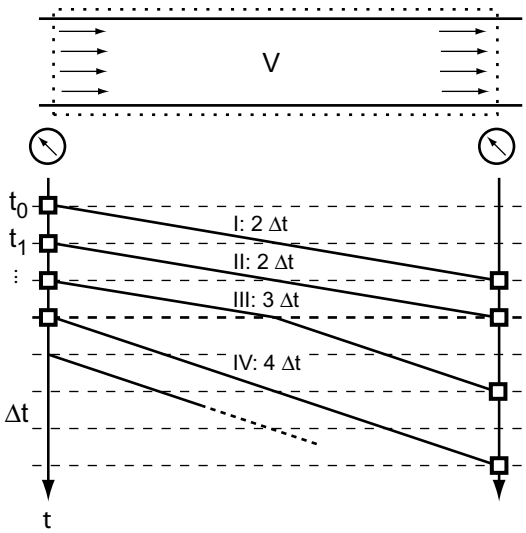

Figure 4.2: Illustration of the time required for water packets I-IV to flow through an ideal plug-flow reactor. The flow rate is unknown, but it can be calculated given the observed travel times and the volume of the reactor.

follows:

$$V = \Delta t Q_0 + \Delta t Q_1 \tag{4.6}$$
$$V = \Delta t Q_1 + \Delta t Q_2 \tag{4.7}$$
$$V = \Delta t Q_2 + \Delta t Q_3 + \Delta t Q_4 \tag{4.8}$$
$$V = \Delta t Q_3 + \Delta t Q_4 + \Delta t Q_5 + \Delta t Q_6 \tag{4.9}$$
$$V = \ldots \tag{4.10}$$

Given enough observations, this equation system can theoretically be solved for the unknown discharges Q. If the response times are relatively short compared to the variation in Q, it is feasible to set $Q_k = \frac{V}{\hat{\theta}_k}$ and, equivalently, to set $\theta_k = \hat{\theta}_k$.

As seen in this relatively simple example of an ideal plug-flow reactor, it is possible to determine the discharge through the reactor given a method which observes the shift of patterns in the time domain. This conclusion is only valid for ideal plug-flow reactors, however. In a tracer experiment, the time that elapses between the addition of a tracer substance to the reactor influent and the statistical *mode* of its response at the effluent corresponds to the $\hat{\theta}$ observed following the rules detailed above. However, the true θ for the tracer experiment is the time span between the tracer addition and the *mean* of the response at the effluent which means that the method is principally only valid for symmetrical residence time distributions; reactor systems that induce dispersion and those in which reactions take place have skewed distributions. The magnitude of the systematic error associated with applying this method to a system with dispersion and reactions is investigated in the next section.

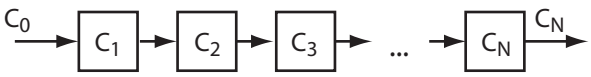

Figure 4.3: A tanks-in-series model consisting of a cascade of N continuous stirred-tank reactors.

4.2.1 Theoretical analysis

The relation between the HRT in the reactor and the tracked travel time for the situation depicted in Figure 4.1 is mathematically modeled by two parallel tanks-in-series models that have a closed-form solution; each of the two models consists of a cascade of N continuous stirred-tank reactors (CSTRs), as shown in Figure 4.3. The closed-form solution simplifies the evaluation of the relation and helps in assessing the extent of the systematic error that was introduced in the section above. It is further assumed that both cascades have the same properties; thus, they are modeled by the same number of reactors N, have the same total volume V and in both cascades there is a first-order degradation reaction, r, taking place. Although this theoretical analysis only holds for compounds for which the mass balance applies, the derivation for temperatures T, thus by formulating a heat balance, is straightforward.

The mass balance of compound C over reactor j in a tanks-in-series model with $1 \leq j \leq N$ equal reactors assuming constant reactor volume $V_j = \frac{V}{N}$, is

$$\frac{dC_j}{dt} = \frac{1}{\theta_j}\left(C_{j-1}(t) - C_j(t)\right) + r_j(t) \tag{4.11}$$

where the HRT in the reactor is denoted by $\theta_j = \frac{V_j}{Q} = \frac{\theta}{N}$. The HRT of the entire cascade is θ, and r_j is the first-order reaction defined by $r_j = -kC_j(t)$ with the reaction constant k.

Let the influent discharge, Q, be constant while the influent concentration, C_0, periodically oscillates according to

$$C_0(t) = a\sin\left(2\pi f t + b\right) + c \tag{4.12}$$

where a is the amplitude, f is the frequency, b is the relative phase shift and c is an offset.

The set of ordinary differential equations $(\frac{dC_1}{dt}, \frac{dC_2}{dt}, ..., \frac{dC_N}{dt})$ that defines the tanks-in-series model has a closed-form solution for the given influent discharge and influent concentration. The asymptotic solution (independent of the initial conditions) for the effluent concentration of reactor N is

$$\begin{aligned} C_N(t) \;=\; & a\left(\frac{1}{\sqrt{(1+k\theta/N)^2+(2\pi f\theta/N)^2}}\right)^N \cdot \\ & \sin\left(2\pi f t + b - N\cdot\arctan\left(\frac{2\pi f\theta}{N+k\theta}\right)\right) + \\ & c\left(\frac{1}{1+k\theta/N}\right)^N \end{aligned} \tag{4.13}$$

Similar to the influent series in Eq. (4.12) (which is, in fact, the special case of $N = 0$), Eq. (4.13) too is a harmonic oscillation. However, if $N > 0$, the amplitude is lower, and when $k > 0$, the offset c decreases (and increases for $k < 0$). In addition, the effluent series exhibits an additional phase shift compared to the influent signal. The difference in the relative phase

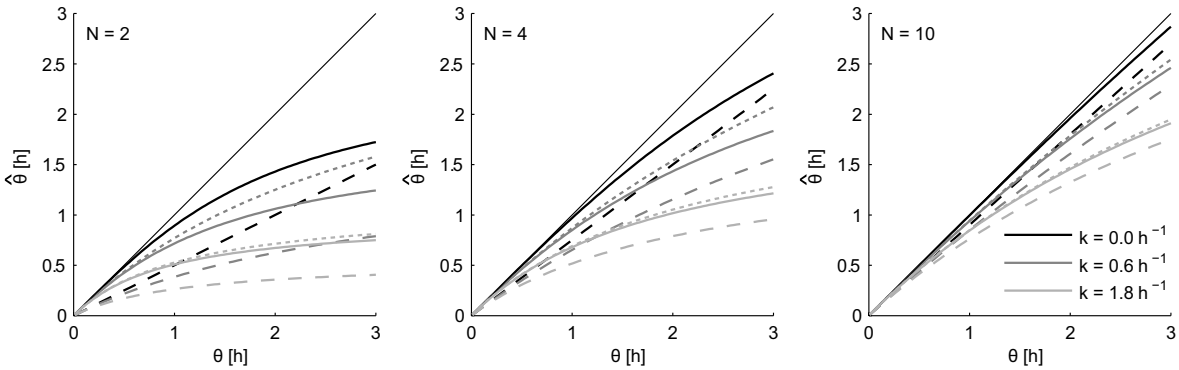

Figure 4.4: The deviation of the observed travel time $\hat{\theta}$ from the HRT θ for nine different tanks-in-series models (full lines) calculated from Eq. (4.14). For each system, the mean (dotted lines) and the mode (dashed lines) of the response of a reactive tracer addition are plotted. The influent function had a frequency of $f = 0.2$ hr^{-1}.

shift between the influent and the effluent signal divided by $2\pi f$ corresponds to the observed travel time and is given by

$$\hat{\theta} = \frac{N}{2\pi f} \arctan\left(\frac{2\pi f \theta}{N + k\theta}\right) \tag{4.14}$$

It is clear that $\hat{\theta} = \theta$ is only valid as $N \to \infty$, in which case the cascade approximates plug-flow behavior (Gujer, 2008). The inversion of the equation system as given in Eqs. (4.6–4.10) is not needed here because of the constant flowrate, Q.

In Figure 4.4, the deviation of $\hat{\theta}$ from θ is plotted for several systems with a finite number of CSTRs. In general, the lower the dispersion in the system (the amount of dispersion decreases with increasing N) and the lower the reaction constant k, the better the conformance.

Systematic error for a single tanks-in-series model

If θ is approximated by $\hat{\theta}$ when $N \ll \infty$ and $k \neq 0$, a systematic error is introduced. The relative error $E_{\theta,rel}$ is defined as

$$E_{\theta,rel} = \frac{1}{\theta}\left(\theta - \hat{\theta}\right) \tag{4.15}$$

and, when applying Eq. (4.14) it is given by

$$E_{\theta,rel} = 1 - \frac{N}{2\pi f \theta} \arctan\left(\frac{2\pi f \theta}{N + k\theta}\right). \tag{4.16}$$

Table 4.1: Relative error of θ and of ξ_A calculated using Eq. (4.16) and (4.19) for typical configurations of different WWTP treatment steps. The volume V and discharge Q_A are given for lane A ($P = 2$); the values for N and k were estimated from measured temperature data of a WWTP with 65,000 people equivalents. The frequency f was set to 0.3 hr^{-1} for the grit chamber and the pipeline and 0.1 hr^{-1} otherwise. $\xi_A = 0.05$ was chosen for the calculation of $E_{\xi_A,rel}$.

System	θ_A [hr]	N [-]	k [hr^{-1}]	$E_{\theta_A,rel}$ [-]	$E_{\xi_A,rel}$ [-]
Grit chamber ($V = 90$ m^3, $Q_A = 0.1$ m^3/s)	0.25	4	0.1	0.01	0.02
Primary clarifier ($V = 750$ m^3, $Q_A = 0.1$ m^3/s)	2	2	0.02	0.12	0.22
Anoxic zone (AST) ($V = 450$ m^3, $Q_A = 0.1$ m^3/s)	1.5	2	≈ 0	0.03	0.06
Pipeline ($V = 35$ m^3, $Q_A = 0.1$ m^3/s)	0.1	12	0	< 0.01	< 0.01

Generally, the error increases with increasing dispersion and increasing reaction rate. The first-order Taylor series approximation $\tan(\varepsilon) \approx \varepsilon$, which is valid for small ε and thus applies for high N, is

$$E_{\theta,rel,\varepsilon} = 1 - \frac{N}{N + k\theta} \approx 0, \tag{4.17}$$

again showing that the error diminishes with decreasing dispersion. Letting $f \to \infty$ and noting that $\lim_{\varphi\to\infty} \tan(\varphi) = \frac{\pi}{2}$, we obtain

$$E_{\theta,rel,\varphi} = 1 - \frac{N}{4f\theta}. \tag{4.18}$$

This result shows that high influent signal frequency increases the systematic error.

To summarize, the systematic error in the approximation of θ with $\hat{\theta}$ is acceptable for systems with low dispersion, a low reaction rate and a lack of high-frequency parts in the influent signal. For comparison, the relative error, $E_{\theta,rel}$ for four different systems commonly found on WWTPs are listed in Table 4.1.

Systematic error for two parallel tanks-in-series models

The discharge distribution is described by the coefficients ξ_A and ξ_B ($\xi_A = -\xi_B$ for $P = 2$), as seen from Eq. (4.5). Let $E_{\xi_A,rel}$ be the systematic error of ξ_A:

$$E_{\xi_A,rel} = 1 - \frac{\hat{\theta}_B - \hat{\theta}_A}{\hat{\theta}_B + \hat{\theta}_A} \frac{\theta_B + \theta_A}{\theta_B - \theta_A} \tag{4.19}$$

The conclusions are the same as for the systematic error of a single tanks-in-series model. The error $E_{\xi_A,rel}$ for common systems is listed in Table 4.1 and is compared to $E_{\theta,rel}$.

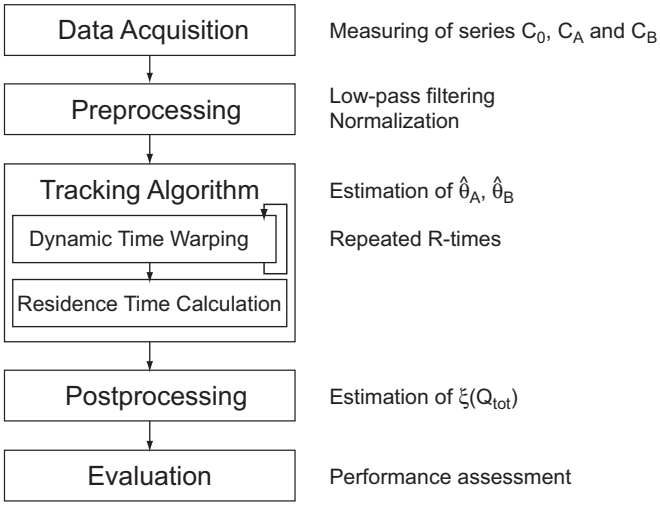

Figure 4.5: Procedure consisting of five successive steps to estimate the discharge distribution at a flow divider.

4.2.2 Procedure

This paper suggests a procedure of five successive steps to estimate the discharge distribution from WWTP measurements (cf. Figure 4.5).

Data requirements and acquisition

The procedure is rather parsimonious in terms of data requirements: Basically, any kind of signal which naturally shows variability can be considered. For the situation illustrated in Figure 4.1, three measurements, C_0, C_A and C_B, are required. Appropriate measuring equipment and location will exhibit the following properties:

- an accurate sensor and a data logger with high resolution, both of which are low in maintenance and inexpensive;
- little (or no) influence of the reactions on the measured variable; and
- variability in the influent and effluent signals.

Temperature and conductivity signals are appropriate choices, among others. The sampling interval, Δt, should be small and the data length L long. For reference, a sampling interval of 30 seconds and a data length of 3 days was appropriate in the case study.

Reasonable choices for the measuring location are the flow divider (influent measurement) and the effluent of each of the branching lanes. Hydraulic residence times and dispersion should be kept to a minimum. This goal can sometimes be achieved by mounting the probe at the opening of a dividing wall or a similar structure instead of the reactor effluent.

Preprocessing

In systems with dispersion, high-frequency noise in the influent is absorbed and is not visible in the effluent. Because high-frequency input signals tend to have high systematic errors (cf. Section 4.2.1) and because they hinder the tracking algorithm that will be introduced in the next section, they must be filtered out.

A finite impulse response *low-pass filter* with a Hamming window (Hamming, 1998) is recommended. A good choice for the cut-off frequency is $f_{cut} = (\theta_{max,est})^{-1}$, which is the inverse of the estimated maximum hydraulic residence time in the reactor. The sensitivity of this parameter on the performance is low, as it will be shown later in Section 4.4.5.

Subsequently, the data is *normalized* to have zero mean and unit variance and down-sampled to a sampling interval of $\Delta t =$ 1-2 minutes. It must be ascertained that all three series have the same time axis.

Tracking algorithm

Dynamic time warping (DTW) is used to find the correspondents of characteristic influent patterns in the effluent signal. DTW assigns each effluent data point a corresponding data point in the influent. DTW is applied twice, once to find $\hat{\theta}_A$ by comparing C_0 with C_A and once to find $\hat{\theta}_B$ by comparing C_0 with C_B. Given $\hat{\theta}_A$ and $\hat{\theta}_B$, ξ_A and ξ_B can be computed using Eq. (4.5).

Basically, the derivation of ξ_A and ξ_B is also possible by comparing the effluents C_A and C_B. However, this derivation requires the total inflow, Q_{tot}, to be known and is therefore not investigated further.

Dynamic Time Warping DTW is a particular implementation of dynamic programming (Bellman, 1957); it is used to optimally align two sequences by non-linearly warping the time-axis of the sequences until their dissimilarity is minimized. In other words, it identifies a warping path that contains information on how to translate, compress and expand patterns so that similar features are matched (Jun, 2011).

Originally, DTW was applied in the field of speech recognition (Sakoe, 1978), but it is now also used in other fields for sequence alignment and as dissimilarity measure.

In contrast to common applications of DTW, not the aligned sequences and the dissimilarity measures are of interest here but rather the warping path itself. The path contains a mapping of all the points of an influent series to the points of an effluent series, which makes it, as stated above, an estimate for the hydraulic residence time.

To align the two sequences $X = (x_1, x_2, \ldots, x_n, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_m, \ldots, y_M)$ with DTW, a distance matrix $\mathbf{D} \in \mathbb{R}^{N \times M}$ with the Euclidian distance between all points of the two series

$$D_{n,m} = \sqrt{(x_n - y_m)^2} \tag{4.20}$$

is first computed.

A warping path, $W = (w_1, w_2, \ldots, w_k, \ldots, w_K)$, is then a sequence of continuous matrix elements that define a mapping between A and B with the k-th element being $w_k = (n, m)_k$. The warping path must satisfy the following conditions (Müller, 2007):

- Boundary condition: The warping path starts at $w_1 = (1, 1)$ and ends at $w_K = (N, M)$ (i.e., the path starts and ends in diagonally opposite corners of the matrix).
- Step size condition: $w_k - w_{k-1} \in \{(0, 1), (1, 0), (1, 1)\}$ (i.e., allowed steps are restricted).

and additionally for this application the constraint:

- Flow direction condition: Given $w_k = (n, m)$ $m \geq n$ (i.e., a pattern appears first in sequence X and then in sequence Y).

As a consequence of the first condition, the algorithm needs some "running-in" time to achieve an appropriate alignment. The affected parts can easily be excluded, however. Please note that the "step size condition" implies that the path is continuous.

While many warping paths that satisfy these conditions exist, the interest lies in the particular path that minimizes the total distance, d, defined by

$$d(X, Y) = \sum_{w_k} D_{n,m}. \tag{4.21}$$

This path can be efficiently found by evaluating the recurrence

$$p_{(n,m)} = D_{n,m} + \min\left(p_{n-1,m-1}, p_{n-1,m}, p_{n,m-1}\right), \tag{4.22}$$

where the cumulative distance, $p_{n,m}$, is defined as the distance in the cell (n, m) and $p_{n-1,m-1}$, $p_{n-1,m}$, and $p_{n,m-1}$ are the minimal cumulative distances of the neighboring cells obtained through dynamic programming (Keogh and Pazzani, 1999). The algorithm is illustrated in Figure 4.6.

Stochasticity DTW does not consider the absolute value of the differences in the distances between neighboring cells; it always finds the path with the minimum cumulative distance. When using noisy and erroneous input signals however, nonrealistic warping paths may result when the distances are below the accuracy of the measuring device.

To generate smoother warping paths, the computation of the warping path is repeated R times. In each run, a random term, $\varepsilon = \mathcal{N}(0, \sigma_m)$, which is normally distributed with zero mean and a standard deviation σ_m set to the accuracy of the measuring device, is added to each data point of the influent and effluent series.

The R individual warping paths are then combined into an averaged warping path $\bar{W}$ by calculating the mode along the diagonal axis $(1, 1) - (M, N)$, as shown in Figure 4.6.

Residence Time Calculation Given the time series C_0 and C_A, each element in the computed averaged $\bar{W}_A$ represents the mapping of the i-th point in time series C_0, $c_{0,i}$ at t_i to the j-th point in series C_A, $c_{A,j}$ at t_j. Recall that the mapping is the result of tracking an imaginary water packet through a reactor. The difference $t_j - t_i$ is thus the travel time of the packet in the reactor. If the travel time is short compared to the variability of Q, then

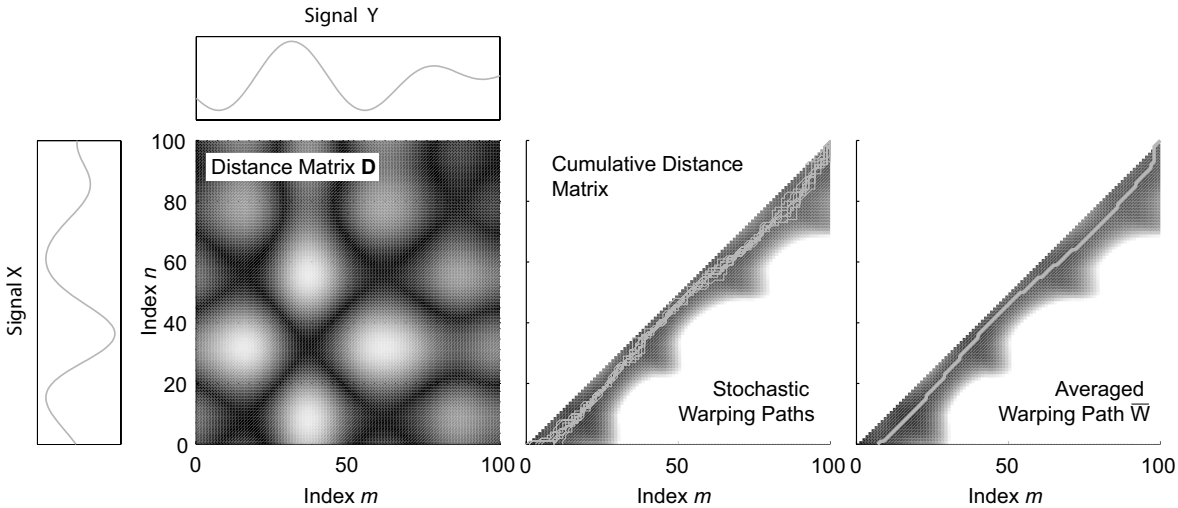

Figure 4.6: Illustration of the DTW algorithm. Given two time series, X and Y, the distance matrix, $\mathbf{D}$, for $m \geq n$ is first calculated (Eq. 4.20). The cumulative distance matrix is computed by solving Eq. (4.22) and is applied to find the warping path, W, with the smallest cumulative distance. This is repeated R times and yields R warping paths, which are averaged to warping path $\bar{W}$.

$\hat{\theta}_A \approx t_j - t_i$. Otherwise, a linear equation system, as given in Eq. (4.6), can be set up and solved. The same procedure applies to the calculation of $\hat{\theta}_B$ using C_0 and C_B.

Special attention must be given to mappings for which $t_j - t_i = 0$. These occur if DTW selects matrix elements for the warping path that lie on the diagonal of the cumulative distance matrix. They are physically not meaningful and must be dropped.

Postprocessing

So far, the procedure has supplied two new series, $\hat{\theta}_A(t)$ and $\hat{\theta}_B(t)$, both containing approximations of the current residence time in lanes A and B at time t.

For further analysis, and to correct the assumption of equal distribution following Eq. (4.2), ξ_A and ξ_B must be computed as a function of Q_{tot}. This is possible using the approximations $\hat{\theta}_A$ and $\hat{\theta}_B$ as follows:

$$Q_{tot}(t) \approx \frac{V}{\hat{\theta}_A(t)} + \frac{V}{\hat{\theta}_B(t)} \tag{4.23}$$

To get a smooth function $\hat{\xi}(Q_{tot})$, a polynomial of degree Z,

$$\hat{\xi}_A(Q_{tot}) = -\hat{\xi}_B(Q_{tot}) = c_0 + c_1 Q_{tot} + c_2 Q_{tot}^2 + \ldots + c_Z Q_{tot}^Z, \tag{4.24}$$

is fitted using the least-squares method. The optimal choice of Z is facilitated by considering an information criterion, such as the Akaike information criterion (AIC), that describes the tradeoff between accuracy and model complexity and helps to prevent over-fitting. The computation of confidence bounds is also recommended because such bounds are valuable for diagnostics.

Table 4.2: Set of parameter values used to build the synthetic systems. For the influent series (a superposition of sine-waves and a noise term), the parameter values for the Ornstein-Uhlenbeck-process generating the noise are specified (specifically the mean value μ_{OU}, the volatility σ_{OU} and the revision rate λ_{OU}).

Parameter	Values
Mean HRT θ_{mean}	15 min, 30 min, 45 min, 60 min
Number of CSTRs in cascade N	2, 3, 5, 10
Reaction rate constant k	0, 0.01 hr^{-1}, 0.03 hr^{-1}, 0.1 hr^{-1}
Inflow discharge series, $Q_{tot}(t)$	
High noise	$\mu_{OU} = 0$ m^3/s, $\sigma_{OU} = 6 \cdot 10^{-4}$ m^3 $\text{s}^{-3/2}$, $\lambda_{OU} = 3 \cdot 10^{-3}$ s^{-1}
Medium noise	$\mu_{OU} = 0$ m^3/s, $\sigma_{OU} = 2 \cdot 10^{-4}$ m^3 $\text{s}^{-3/2}$, $\lambda_{OU} = 3 \cdot 10^{-4}$ s^{-1}
Low noise	$\mu_{OU} = 0$ m^3/s, $\sigma_{OU} = 7 \cdot 10^{-6}$ m^3 $\text{s}^{-3/2}$, $\lambda_{OU} = 3 \cdot 10^{-5}$ s^{-1}
Inflow temperature series, $T_0(t)$	
High noise	$\mu_{OU} = 0$ °C, $\sigma_{OU} = 7 \cdot 10^{-3}$ °C $\text{s}^{-1/2}$, $\lambda_{OU} = 3 \cdot 10^{-3}$ s^{-1}
Medium noise	$\mu_{OU} = 0$ °C, $\sigma_{OU} = 2 \cdot 10^{-3}$ °C $\text{s}^{-1/2}$, $\lambda_{OU} = 2 \cdot 10^{-4}$ s^{-1}
Low noise	$\mu_{OU} = 0$ °C, $\sigma_{OU} = 7 \cdot 10^{-4}$ °C $\text{s}^{-1/2}$, $\lambda_{OU} = 4 \cdot 10^{-5}$ s^{-1}
Discharge distribution	$\xi = 0.22$

Model evaluation

Several tests that give insights into the performance of the model can be carried out. If any of the following tests fail, the application of the procedure was not successful or there was no clear relationship between discharge distribution and total inflow.

- If a measurement of Q_{tot} is available, it can be compared to the approximation calculated with Eq. (4.23). High deviations indicate bad performance.
- If many mappings had to be dropped because $t_j - t_i = 0$, there were difficulties finding the warping path.
- If the polynomial $\hat{\xi}(Q_{tot})$ has wide confidence bounds, there was no clear trend in the data.

4.3 Software availability

A MATLAB implementation of the procedure, together with example data, is available (cf. Appendix A.2, page 96).

4.4 Results and discussion

4.4.1 Synthetic systems

The performance of the proposed method was assessed by calculating the discharge distribution for synthetic systems consisting of two tanks-in-series models in parallel. The number of reactors in the models, the mean hydraulic residence times and the reaction constant of the first order degradation reaction were varied (cf. Table 4.2).

Realistic synthetic discharge and influent temperature series were generated. They exhibited three different levels of variability that corresponded to different measuring locations within

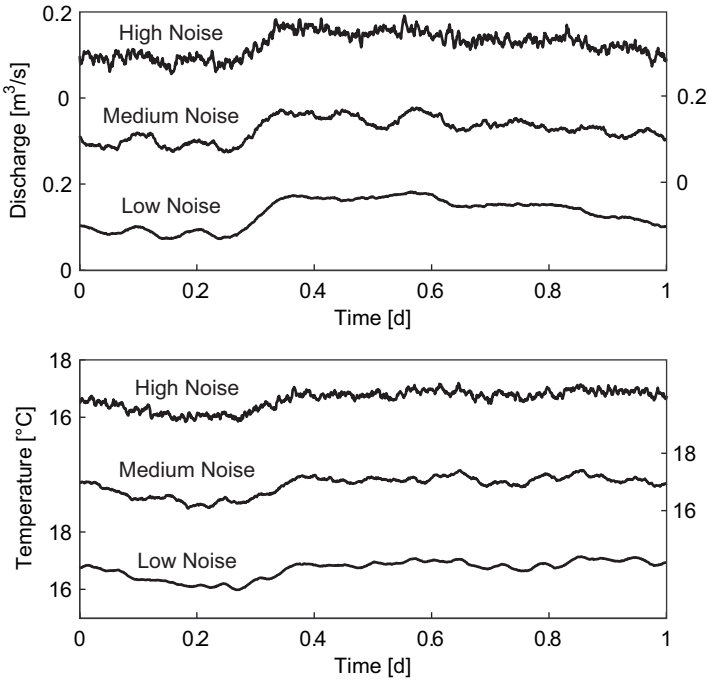

Figure 4.7: One day of the synthetic influent discharge and temperature series used as input for the synthetic systems. The parameters for the series are given in Table 4.3.

Table 4.3: The parameter values used for the calculations presented for the synthetic systems and the system of the case study.

Parameter	Synthetic System	Grit Chamber
Δt	2 min	2 min
L	4 days	3 days
f_{cut}	Q_{min}/V	1/0.75 hr^{-1}
σ_m	0.25 °C	0.25 °C
R	500	500
Z	5	5

the flow scheme of a WWTP. They consisted of a combination of superimposed sine waves and a stochastic component generated by an Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930). The different influent series are plotted in Figure 4.7, and the parameter values are given in Table 4.2. The discharge distribution was predefined; the effluent temperature for each of the two models was computed by solving the models with a Runge-Kutta (4,5)-solver (Shampine, 1994).

With the generated influent temperature series and the two modeled effluent temperature series, the procedure could be applied. The values of the parameters for the procedure are listed in Table 4.3.

To assess the performance, the coefficient of variation of the root mean squared deviation, CV(RMSD), Eq. (4.25), was calculated to measure the deviation of the estimated discharge distribution $\hat{\xi}_A(Q_{tot})$ from the pre-defined distribution (cf. Table 4.2) and to indicate how well the discharge Q_A corresponds to the pre-defined discharge function.

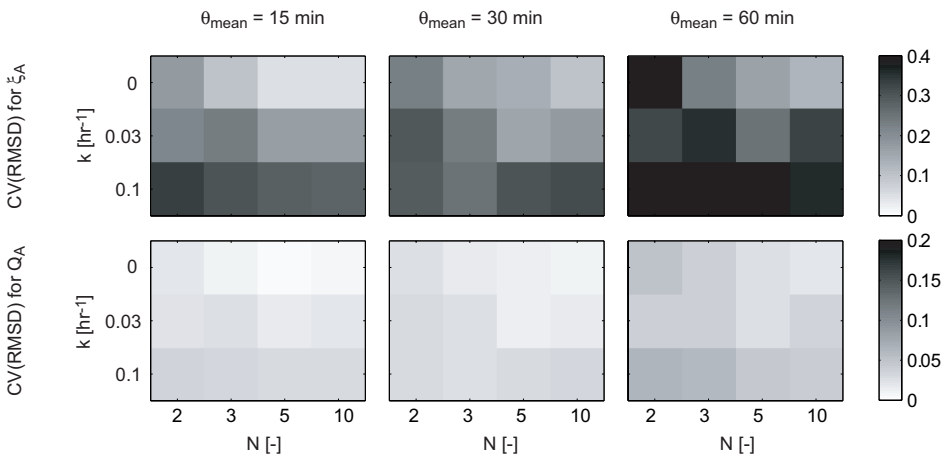

Figure 4.8: CV(RMSD) for the calculation of $\xi(Q_{tot})$ (upper row) and the corrected discharge $Q_A(t)$ (lower row) for three different mean hydraulic residence times θ_{mean} (columns) with varying N and k.

$$\mathrm{CV(RMSD)} = \frac{\mathrm{RMSD}}{\bar{x}} \tag{4.25}$$

with

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{1,i} - x_{2,i}\right)^2}{n}}. \tag{4.26}$$

After computing 345 different systems constructed using parameter values from the sets defined in Table 4.2, the influence of the parameter set on the resulting fit was investigated with the ReliefF measure. The ReliefF measure is a measure for attribute weighting (Robnik-Sikonja and Kononenko, 2003). This analysis revealed that for the deviation of both $\hat{\xi}_A$ and Q_A from their predefined values, the parameters can be grouped into three sets, sorted by descending influence on the fit: $\{k\}$, $\{N, \theta_{mean}\}$ and $\{Q_{in}(t), T_0(t)\}$. This result implies that the level of noise in the time series $Q_{in}(t)$ and $T_0(t)$ has little effect on the fit and that the reaction coefficient, k, is the most decisive parameter. It is thus important to carefully check that there is no significant reaction taking place that may influence the signal. The matrices shown in Figure 4.8 compare the CV(RMSD) for different combinations of N, k and θ_{mean}. The performance measures of all synthetic systems are available in the supporting information.

Comparing the corrected discharge $Q_A(t)$ and the uncorrected discharge $Q_a(t)$ (which assumes that the flow is equally split) with the pre-defined flow, it is clear that the correction drastically improves the fit: While the uncorrected discharge Q_a leads to an error of CV(RMSD) = 0.11, the corrected series Q_A has an error between CV(RMSD) = 0.006 and 0.06.

4.4.2 Case study: grit chamber and primary clarifier

The practical applicability of the procedure was tested in a mid-sized Swiss WWTP (65,000 people equivalents). Between the fine screen and the grit chamber, there is a hydraulic flow divider that splits the incoming discharge into two branches, each of which is fed to an aerated grit chamber followed by a primary clarifier. After the latter, the flows join again.

Although the flow divider was built to provide a uniform distribution, the WWTP staff suspected it to be irregular. However, because there is no discharge measurement in either of the two lanes (only the combined flow is known), the magnitude of the inequality could not be assessed.

The proposed procedure was used to quantify the discharge distribution. The temperature in the flow divider and in the effluents of the grit chambers (by design $\theta_{mean} = 15$ min) was measured using TMC6-HD temperature probes with an accuracy of 0.25 °C in combination with U12-006 temperature loggers with a data resolution of 0.03 °C (both manufactured by Onset, Bourne, MA, United States). The temperature measurement were chosen because it can easily and accurately be measured with inexpensive autonomous devices. In addition, there are no significant reactions that affect the wastewater temperature.

The parameter values for the procedure are given in Table 4.3. The parameter σ_m was set to the accuracy of the temperature probe, and f_{cut} was set to the inverse of an estimate of the maximal hydraulic residence time in the reactor.

To validate the results, the discharge in each branch was measured within the ducts connecting the grit chambers and the primary clarifiers using a mobile Venturi flume (Hager, 1999).

The measured temperature series and the measured and estimated discharge distribution function $\xi(Q_{tot})$ are plotted in Figure 4.9. In addition, the measured discharge is compared to the discharge Q_a, which assumes a uniform distribution, and the discharge Q_A which is corrected using Eq. (4.2).

It is visually clear that the discharge is not uniformly distributed; there is always a higher flow in branch A, especially at low and high total discharges. To guarantee efficient grit removal, the given grit chamber is designed to have an HRT of 15 minutes at a discharge of $Q_{tot} = 0.2$ m^3/s ($V = 90$ m^3 per lane). Assuming that for $\xi_A = 0.2$ the HRT is reduced to 12.5 minutes, performance losses are possible at high influent discharges.

Considering the tasks enumerated in the Introduction, the operator on the one hand can now, iteratively, adjust a gate valve upstream of the grit chamber or perform other constructional modifications and reassess the situation until the unevenness is no longer significant.

The scientist or engineer on the other hand now possesses a function that he or she can use to obtain a more appropriate discharge series for reactor modeling. Assuming uniform distribution, the CV(RMSD) of the discharge series is 0.14; when using the corrected discharge series instead, the error is reduced to 0.06.

4.4.3 Validation

There are several means of judging the quality of the outcome when the given procedure is applied. Three quick tests to check whether the procedure has successfully found the function $\hat{\xi}(Q_{tot})$ have already been described in Section 4.2.2. In addition to these, the distribution of

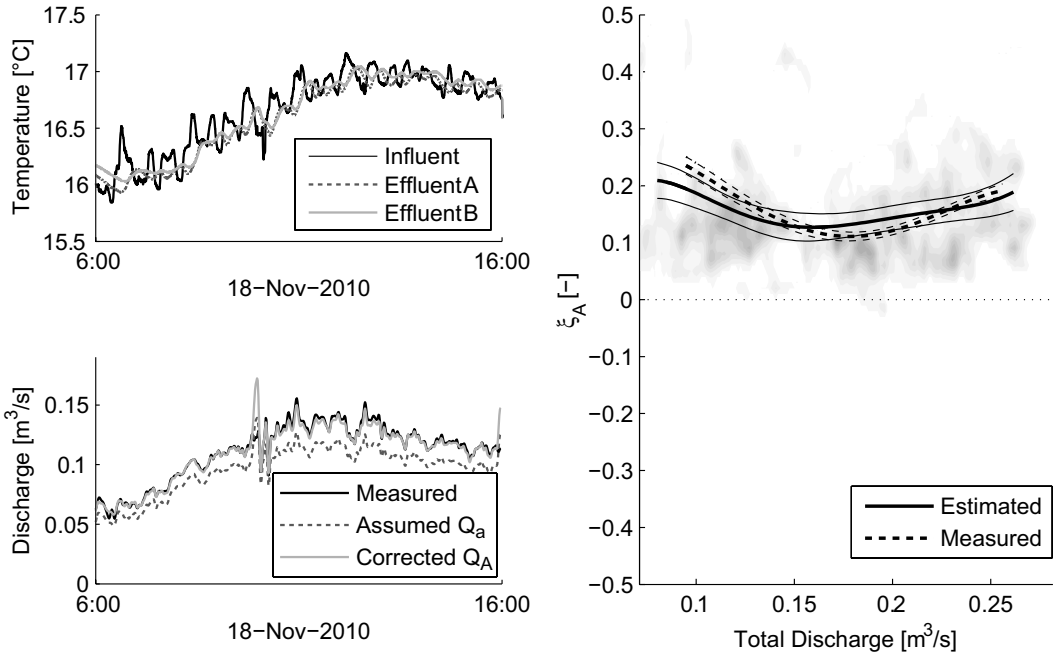

Figure 4.9: Determination of the discharge distribution at a grit chamber operated in two parallel lanes. Temperature measurements at the flow divider and the lane effluents (top left); estimated (procedure) and measured (mobile Venturi flume) discharge distribution with 99% confidence bounds (right); comparison of the measured discharge (mobile Venturi flume), the discharge assuming uniform distribution (based on WWTP total influent measurements) and the corrected discharge (bottom left). Please note that on the left, only a section of the validation time series are plotted (total length: three days).

the individual points $\xi(Q_{tot})$, as shown in Figure 4.9 (right) as a point density cloud, gives further insight into the quality of the outcome and is discussed below.

A cloud showing a clear trend and a fitted polynomial with narrow confidence bounds indicate success. A cloud that is scattered and almost has the form of a circle, however, may indicate inappropriately chosen parameter values (data length L too short, Δt too high), significant reactions that influence the measured signal, or high dispersion. The latter can sometimes be circumvented by choosing a better measuring location.

A forking cloud may indicate two different "regimes". Intermittent lateral inflows that are added before the flow divider and that influence the flow are a possible reason for this. It is sometimes possible to disconnect these flows during the experiment. Otherwise, if the times when lateral inflows occur are known, the corresponding sequences can be removed from the measured time series and the individual windows can be analyzed.

If these suggestions do not help and the quality of the outcome remains unclear, it is advisable to consider an alternative approach. However, it is worth mentioning that there is little risk that the discharge $Q_i(t)$, calculated by Eq. (4.2), is a worse estimate than the flow $Q_a(t)$ under the assumption of equal distribution. From the theoretical analysis, $\hat{\theta}_A$ and $\hat{\theta}_B$ are underestimated in case of dispersion and reaction ($k > 0$) and $\left|\hat{\theta}_A - \hat{\theta}_B\right| < |\theta_A - \theta_B|$; therefore the estimates of ξ_A and ξ_B using $\hat{\theta}_A$ and $\hat{\theta}_B$ following Eq. (4.5) approach zero for increasing reaction and dispersion, and thus $Q_i(t)$ goes towards Q_a(t).

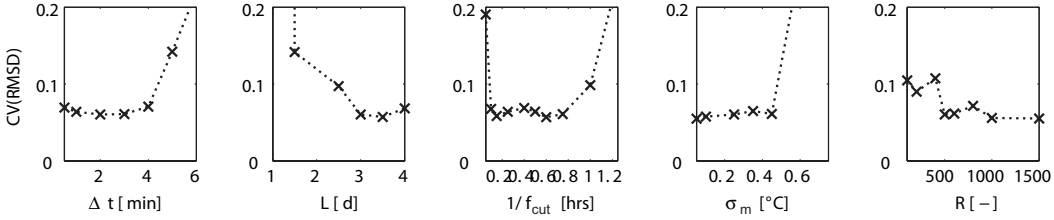

Figure 4.10: Effect of the parameters for the procedure on the fit of Q_A, calculated for the "case study" experiment.

4.4.4 Optimal input data

The use of temperature measurements for quantifying discharge distribution has proven to be a sound choice. Autonomous, highly accurate temperature probes with high-resolution loggers are readily available. Temperature usually exhibits sufficient variability and is often barely affected by reactions (cf. Table 4.1). In aerated basins, however, the temperature may change substantially, and an alternative must be found.

Conductivity measurements are a possible alternative that the authors have evaluated. They have the following deficiencies, however:

- conductivity sensors are rather maintenance-intensive and costly; and
- there are large conductivity changes due to biological reactions, particularly in the activated sludge tanks.

Nevertheless, conductivity measurements exhibit good variability and may be considered for the calculation of the discharge distribution when the use of temperature measurements fail.

4.4.5 Choice of parameter values for the procedure

The effect of the choice of parameter values for the procedure on the goodness of the discharge Q_A was investigated for the "case study" experiment. For this purpose, one parameter was varied at a time within a range while the others were left unchanged, and the fit of the estimated ξ_A was calculated. The results are illustrated in Figure 4.10.

It is clear that the fit shows little sensitivity to the choice of parameter values for Δt, f_{cut} and σ_m. The length of the measured series L should be at least three days. For the number of runs R, the variability of the error is very high for a low R but converges as the value or R increases.

4.4.6 Further applications

The application of the procedure introduced in this paper is not limited to the assessment of discharge distribution at hydraulic flow dividers. Estimation of the flow through a reactor and the estimation of the flow velocity in a channel were not discussed but are also within the range of possible applications. These estimates could be used for the diagnosis of flow measurement devices. In a scenario in which the discharge is known, the delay of the characteristic patterns may provide insight into the mixing processes and reactions.

4.5 Conclusions

In this study, a new method was presented that allows estimation of the discharge distribution at flow dividers used on WWTPs.

- The method requires only measurement of a signal at the flow divider and at the effluent of each reactor downstream of the divider. The signal must have some variability, and it should not be heavily affected by the reactions taking place in the reactors. Temperature sensors have proven to be a good choice because they are robust, commonly available, almost maintenance-free and they deliver high accuracy.
- The method uses an implementation of the dynamic time warping algorithm to assign patterns found in an influent signal to patterns in each of the effluent signals. The time passed between the observation in the influent and in each of the effluents is an estimate for the travel time of a "water packet" in the reactor and relates to its hydraulic residence times. Given these estimates, the discharge distribution between the branches of the dividing structure can be assessed.
- All the parameters for the method can be linked to properties that are generally known or can easily be estimated.
- It was shown that the method is accurate for reactors that have low dispersion and in which reactions do not significantly influence the measured signals.
- Synthetic systems and a real flow divider were analyzed and the discharge distribution was estimated. The relative error in the inflow discharge to a branch of the distribution device can be reduced by more than 50% in comparison to the the assumption of equal distribution by estimating the discharge distribution.

4.6 Supporting Information

- MATLAB source code
- Excel table with the configurations of the synthetic systems and the associated performance indicators

References

Ahnert, M., Kuehn, V., Krebs, P., 2010. Temperature as an alternative tracer for the determination of the mixing characteristics in wastewater treatment plants. Water Research 44 (6), 1765–1776.

Bellman, R., 1957. Dynamic programming. Princeton Univ. Pr, Princeton, NJ.

Dutta, S., Catano, Y., Liu, X., Garcia, M. H., 2010. Computational Fluid Dynamics (CFD) modeling of flow into the aerated grit chamber of the MWRD's north side water reclamation plant, Illinois. World Environ. Water Resour. Congr.: Chall. Change - Proc. World Environ. Water Resour. Congr., 1239–1249.

Gresch, M., Braun, D., Gujer, W., 2010. The role of the flow pattern in wastewater aeration tanks. Water Science and Technology 61 (2), 407–414.

Gujer, W., 2008. Systems analysis for water technology. Springer, Berlin.

Hager, W. H., 1999. Wastewater hydraulics: Theory and practice. Springer, Berlin.

Hamming, R. W., 1998. Digital Filters, 3rd Edition. Courier Dover Publications.

Harremoes, P., Capodaglio, A. G., Hellstrom, B. G., Henze, M., Jensen, K. N., Lynggaard-Jensen, A., Otterpohl, R., Soeberg, H., 1993. Wastewater treatment plants under transient loading- Performance, modelling and control. Water Science and Technology 27 (12), 71–115.

Jun, B. H., 2011. Fault detection using dynamic time warping (DTW) algorithm and discriminant analysis for swine wastewater treatment. Journal of Hazardous Materials 185 (1), 262–268.

Keogh, E. J., Pazzani, M. J., 1999. Scaling up dynamic time warping to massive dataset. Principles of Data Mining and Knowledge Discovery 1704, 1–11.

Müller, M., 2007. Dynamic Time Warping. In: Information Retrieval for Music and Motion. Springer Berlin Heidelberg, pp. 69–84.

Orhon, D., Soybay, S., Tünay, O., Artan, N., 1989. The Effect of Reactor Hydraulics on the Performance of Activated-Sludge Systems - 1. The Traditional Modelling Approach. Water Research 23 (12), 1511–1518.

Patel, T., O'Luanaigh, N., Gill, L. W., 2008. The efficiency of gravity distribution devices for on-site wastewater treatment systems. Water Science & Technology 58 (2), 459.

Port, E., 1994. Anforderungen an die Eigenüberwachung bei kommunalen Kläranlagen. Vol. 75 of Schriftenreihe WAR. WAR Darmstadt, Darmstadt.

Quevauviller, P., Thomas, O., van Der Beken, A., 2006. Wastewater quality monitoring and treatment. Water quality measurements series. Wiley, Chichester.

Robnik-Sikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 53 (1-2), 23–69.

Sakoe, H., 1978. Dynamic-Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics Speech and Signal Processing 26 (1), 43–49.

Shampine, L. F., 1994. Numerical solution of ordinary differential equations. Mathematics. Chapman & Hall, New York.

Uhlenbeck, G. E., Ornstein, L. S., 1930. On the theory of the Brownian motion. Physical Review 36 (5), 823–841.

Chapter 5

Automatic reactor model synthesis with genetic programming

Accepted for publication in Water Science and Technology

David J. Dürrenmatt and Willi Gujer

Automatic reactor model synthesis with genetic programming

David J. Dürrenmatt and Willi Gujer**

* *Institute of Environmental Engineering, ETH Zurich, 8093 Zurich, Switzerland and Swiss Federal Institute of Aquatic Science and Technology, Eawag, 8600 Dübendorf, Switzerland (E-mail: david.duerrenmatt@eawag.ch)*

Abstract Successful modeling of wastewater treatment plant (WWTP) processes requires an accurate description of the plant hydraulics. Common methods such as tracer experiments are difficult and costly and thus have limited applicability in practice; engineers are often forced to rely on their experience only. An implementation of grammar-based genetic programming with an encoding to represent hydraulic reactor models as program trees should fill this gap: The encoding enables the algorithm to construct arbitrary reactor models compatible with common software used for WWTP modeling by linking building blocks, such as continuous stirred-tank reactors. Discharge measurements and influent and effluent concentrations are the only required inputs. As shown in a synthetic example, the technique can be used to identify a set of reactor models that perform equally well. Instead of being guided by experience, the most suitable model can now be chosen by the engineer from the set. In a second example, temperature measurements at the influent and effluent of a primary clarifier are used to generate a reactor model. A virtual tracer experiment performed on the reactor model has good agreement with a tracer experiment performed on-site.

Keywords hydraulic reactor systems; modeling; operating data; grammar-based genetic programming

5.1 Introduction

A key element for successful wastewater treatment plant (WWTP) modeling is an accurate description of the hydraulic processes. In water technology, transport and mixing phenomena can often be approximated sufficiently by cascading ideal reactors such as the continuous-stirred tank reactor (CSTR) and plug-flow reactor (PFR) models (Alex *et al.*, 2002; Gujer, 2008).

The construction of an appropriate model given influent and effluent observations is a widely studied system identification problem and many techniques exist (see, e.g., Ljung, 1987; Keesman, 2011). However, system identification remains a difficult task particularly when the structure of the system is unknown (Flores and Graff, 2005). In systems analysis for water technology, two methods for model identification prevail (Gujer, 2008): The first method consists of the analysis of an experimentally determined impulse response (e.g., by a tracer

experiment) that gives insights into mixing phenomena and that can, for some few ideal reactor systems, also give insights into model structure (Keesman and Stigter, 2002; Gujer, 2008). From a practical point of view, however, tracer experiments are labor-intensive and are hindered by high flow variability, required mixing lengths and density effects; therefore they are often omitted.

The second method consists of the process of manually adjusting model structure and parameters until a pre-defined objective function is minimized. Although this procedure is widely accepted, it bears some problems. Still, extensive measuring campaigns may be required and manual search by trial and error is inefficient as it only covers a small part of the model and parameter spaces, thus the most appropriate model is not guaranteed to be found.

As a consequence, engineers often rely on experience and intuition only, which is dissatisfying because inappropriate hydraulic modeling can have a significant impact on predicted effluent concentrations (Tchobanoglous, 2003). In this paper, a highly practical method is presented that provides the engineer a small set of hydraulic reactor models that perform equally well; the most suitable model can then be selected by taking expert knowledge into account. The set of models is generated using genetic programming (GP), an advancement of genetic algorithms.

GP is an evolutionary computational technique used for optimization and particularly suited for complex problems with high-dimensional search spaces (Koza, 1992) and has previously been considered for system identification, mainly for symbolic regression of sets of ordinary differential equations (for an overview, see Flores and Graff, 2005). For the implementation presented in this paper, in contrast, a special GP tree encoding was defined that encodes hydraulic reactor models. The encoded models consist of the building blocks CSTR and PFR and are therefore compatible with common software used for WWTP modeling. As input, the method requires easily measured signals only (e.g., temperature data), even if they are influenced by physical and chemical reactions, in addition to discharge data.

Two alternative approaches for the direct generation of reactor models worth mentioning have used genetic algorithms (Laquerbe *et al.*, 2001) or found a model by simplifying initially complex super-structures (Hocine *et al.*, 2008). Both approaches, however, require the measurement of the residence time distribution.

5.2 Material and methods

5.2.1 Reactor modeling

The reactor types commonly used to model hydraulic characteristics are the CSTR and PFR. By linking these basic reactors serially or in parallel, even complex situations can be modeled. Each reactor has one or more in- and outflows. (A reactor with more than one outflow can be imagined as reactor with a subsequent flow divider.)

In this study, only one compound is considered and its concentration is calculated for each reactor. The relevant elements for the construction of arbitrary reactor models are the inflow and outflow nodes, CSTRs, PFRs, reactions and fluxes. An example is given in Figure 5.1.

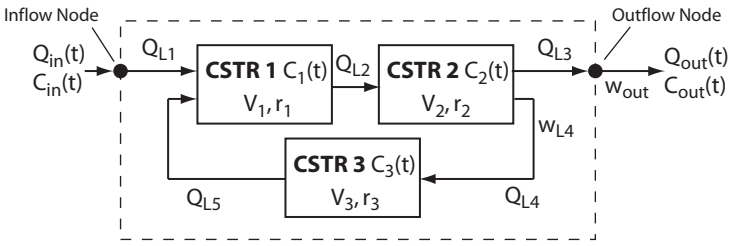

Figure 5.1: An example system consisting of three connected CSTRs.

Inflow and outflow node

Each hydraulic reactor system has an inflow node that acts as a junction node to combine the system influent (with flow Q_{in} and concentration C_{in}) with zero or more return flows. Similarly, the outflow node collects fluxes and releases them from the system. Between the inflow and outflow nodes, there is any number of arbitrarily linked reactors.

CSTRs and PFRs

The CSTR, probably the most important building block, is implemented as an ordinary differential equation (ODE). It is assumed that its volume V is constant over time; the mass balance for compound C in the reactor can therefore be written as

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \frac{1}{V}\left(\sum_j Q_j C_j - C\sum_j Q_j\right) + r \tag{5.1}$$

with one or more inflow Q_j and a reaction term r. PFRs are implemented as a delay without reaction (since delays are typically very small), and the effluent concentration at time t is

$$C_t = C\left(t - \frac{V}{\sum_j Q_j}\right). \tag{5.2}$$

Reactions

Reactions taking place in the CSTRs are modeled with zero-order ($r = k$ with reaction coefficient k), first-order ($r = kC$ with reaction coefficient k) or Monod kinetics ($r = q_{\max}C/(K_{\text{S}} + C)$) with maximal activity $q_{\max}$ and half-saturation coefficient K_{S}. Reactions taking place in the PFRs are not modeled. However, approximating the PFR by a cascade of n CSTRs is an option because plug-flow behavior is approximated for $n \to \infty$.

Fluxes

Because the reactor volumes are constant, the flow in each link can be calculated by solving a linear equation system given the total inflow to the system and, if there are any reactors with more than one effluent, given their flow ratios expressed by the quotient of the weight factors assigned to each of the links. The system depicted in Figure 5.1, for instance, has unknown flows Q_{L1}, Q_{L2}, …, Q_{L5} while the total inflow $Q_{\text{in}}(t)$ as well as the two weight factors w_{L4}

and w_{out} are given. This allows writing the following linear equation system consisting of four mass balance equations

$$Q_{\text{L1}} = Q_{\text{in}} \tag{5.3}$$
$$Q_{\text{L1}} + Q_{\text{L5}} = Q_{\text{L2}} \tag{5.4}$$
$$Q_{\text{L2}} = Q_{\text{L3}} + Q_{\text{L4}} \tag{5.5}$$
$$Q_{\text{L4}} = Q_{\text{L5}} \tag{5.6}$$

and one equation that expresses the ratio of the effluent flows of reactor "CSTR 2"

$$Q_{\text{L3}}/Q_{\text{L4}} = w_{\text{out}}/w_{\text{L4}} \tag{5.7}$$

5.2.2 Genetic programming

GP is a search algorithm inspired by nature (Koza, 1992). It aims to evolve mathematical expressions or computer programs by mimicking biological evolution. Starting with a population of individuals (random programs), new generations are bred. During each generation, the fitness of every individual is evaluated by a fitness function. The fittest ones among the population are then more likely to survive to the next generation. They can be copied unaltered (reproduction), they can feature random changes (mutation), or they can be used to generate new offspring by combining two parents (crossover). This process is repeated until either a given fitness criterion is met or a maximum number of generations is reached.

The evolved, tree-like computer programs can vary in length, which is a valuable characteristic of GP. Other advantages include the absence of a tendency for the entire population to converge and the fact that the form of the solution does not need to be known in advance (Tsakonas, 2006). A function set and a terminal set from which GP can choose to build the programs, in addition to a measure of fitness, need to be specified. In this paper, a grammar-based paradigm is selected with a context-free grammar that consists of a set of terminal nodes, function nodes, a set of reproduction rules that define for each function the possible child function(s) and a starting symbol (the root of the tree). The definition of a grammar avoids the generation of meaningless programs and thereby significantly reduces the search space.

5.2.3 Tree encoding

A tree encoding was defined to represent hydraulic reactor models as computer programs. The functions and terminals available for the program are listed in 5.1, and the grammar rules are given in Table 5.2. Every program starts with a ROOT function, which has two child nodes, that splits the program into two branches. The left branch (starting from the ADF_L function) encodes the layout of the model (the reactors and their connections), whereas the right branch encodes the reaction rate, which can be referenced by some of the reactors of the model. The program tree is recursively decoded starting from the outermost terminals. Once the decoding reaches the ROOT function, the inflow and outflow nodes are added, resulting in an object-oriented representation of the hydraulic model. The decoding is illustrated in Figure 5.2.

Table 5.1: The function and terminal sets for encoding the reactor models.

Name	Description	Num. children (type)
Function set		
ROOT	Root node (starting node)	2 (Function)
ADF_L	Automatically defined function encoding the model layout	1 (Function)
ADF_R	Automatically defined function encoding the reaction	1 (Function)
PAR	Parallel arrangement	2 (Function)
SER	Serial arrangement	2 (Function)
INV	Invert fluxes of child nodes	1 (Function)
CSTR	Continuous stirred-tank reactor	4 (Terminal)
PFR	Ideal plug flow reactor	3 (Terminal)
VLR	Volume-less reactor, used to encode shortcut flows	1 (Terminal)
R_ZERO	Zero-order reaction kinetics	1 (Terminal)
R_FIRST	First-order reaction kinetics	1 (Terminal)
R_MONOD	Mixed-order reaction kinetics (Monod)	2 (Terminal)
Terminal set		
REACTION	Either NO_R (no reaction) or ADF_R (reaction defined in ADF_R branch)	
ERC_vol	Reactor volume (set of constants)	
ERC_flow	Weight factor for flow distribution (set of constants)	
ERC_k0	Zero-order reaction constant (set of constants)	
ERC_k1	First-order reaction constant (set of constants)	
ERC_q	Maximum activity, Monod kinetics (set of constants)	
ERC_KS	Half-saturation coefficient, Monod kinetics (set of constants)	

Table 5.2: Grammar rules to ensure meaningfulness of the evolved computer programs.

Function	Sets of children available for each descendant of the function
ROOT	{ADF_L}, {ADF_R}
ADF_L	{PAR, SER, CSTR, PFR, VLR}
ADF_R	{R_ZERO, R_FIRST, R_MONOD}
PAR	{PAR, SER, INV, CSTR, PFR, VLR}, {PAR, SER, INV, CSTR, PFR, VLR}
SER	{PAR, SER, CSTR, PFR, VLR}, {PAR, SER, CSTR, PFR, VLR}
INV	{PAR, SER, CSTR, PFR, VLR}
CSTR	{ERC_vol}, {ERC_flow}, {ERC_flow}, {REACTION}
PFR	{ERC_vol}, {ERC_flow}, {ERC_flow}
VLR	{ERC_flow}
R_ZERO	{ERC_k0}
R_FIRST	{ERC_k1}
R_MONOD	{ERC_mu}, {ERC_KS}

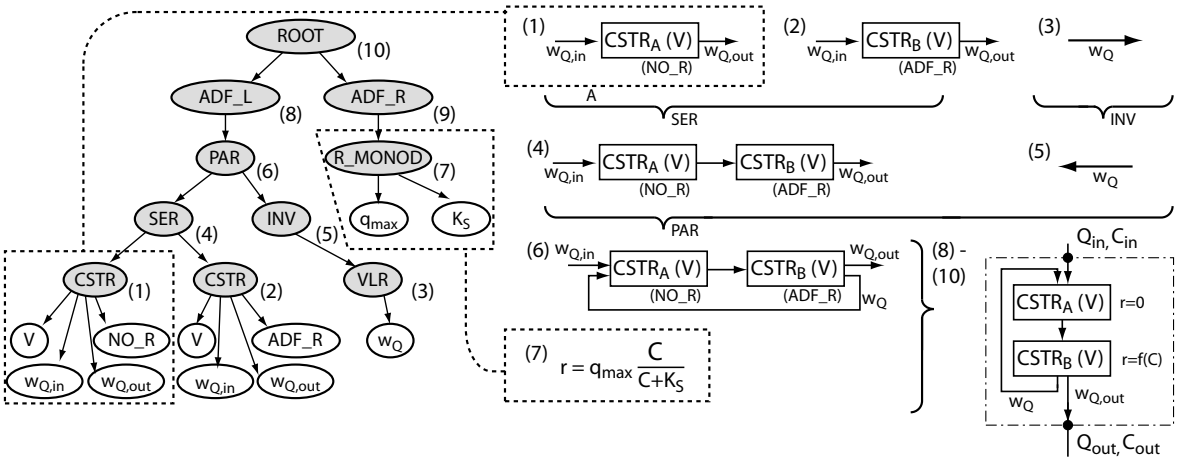

Figure 5.2: An illustration of the decoding of a GP program tree. The tree (left) is recursively traversed from the outermost nodes in ten steps and the reactor model is successively extended. The extension is graphically illustrated (right) and the resulting model shown (bottom right). The volumes V are taken from the ERC_vol set, the weight factors w_{Q} from the ERC_FLOW set and the reaction coefficients $q_{\max}$ and K_{S} from the ERC_q and ERC_KS sets, respectively.

Fitness function

The fitness F of a program tree is a scalar value and is evaluated by a fitness function that takes into account the error between the actual and predicted effluent and the complexity of the model. The consideration of model complexity is important because tree sizes tend to grow during evolution (Koza, 1992; Soule and Foster, 1998), consequently leading to more complex models that have limited generalization ability, i.e. that cause over-fitting of the data (McKay *et al.*, 1997). However, over-fitting can be prevented by introducing a term for "parsimony pressure" that penalizes large tree structures (Soule and Foster, 1998). Thus, the fitness value for an individual is calculated as

$$F = CV(RMSD) + \alpha L \tag{5.8}$$

where CV(RMSD), the coefficient of variation of the root mean square deviation, expresses the error when comparing the prediction of the numerically solved model with the measured series, α is the coefficient for parsimony pressure and L the number of links in the model. Because lower residuals represent better models, the goal is to minimize the value calculated by the fitness function. If a reactor model cannot be solved, an infinite fitness value is returned. This discourages its survival into the next generation.

Selection, crossover and mutation

During evolution, the fittest individuals of each generation are selected for the next generation in tournaments. In each tournament, a number of individuals are picked at random. The best is chosen with selection probability p, the second best with probability $p(1-p)$, etc. Tournament selection allows an easy adjustment of the selective pressure by the tournament size parameter. Some of the selected individuals experience random mutations, where nodes

of the program tree are randomly exchanged. Others are affected by random crossovers, where two individuals exchange parts of their programs. The selection probability and the rates of crossover, mutation and reproduction remain constant during breeding to avoid the convergence of the population and thus getting stuck in a local optimum. The maximum tree depth is constrained to avoid the generation of overly complex models and is therefore another means to prevent over-fitting, next to parsimony pressure introduced in Eq. (5.8).

Procedure

The search algorithm is run several times in parallel to obtain a palette of equally well-performing models. Theoretically, the global optimum will always be found. However, breeding is stopped when a pre-defined fitness criterion is satisfied or when the maximum number of generations has been reached. In addition, due to the noisy nature of measurements, the use of different input data for the individual runs is encouraged. Eventually, the model palette can then be presented to the expert, who selects the most appropriate model.

5.3 Results and discussion

5.3.1 Synthetic system: CSTRs in series with and without reaction

The power of the proposed method was first assessed by investigating its ability to identify a predefined synthetic reactor model in two experiments. The reactor model consisted of two CSTRs in series (volumes of 400 m^3 and 1600 m^3). In the first experiment, no reaction took place in any of the two reactors, whereas in the second experiment, a degradation reaction occurred in the second reactor. The reaction had first-order kinetics with $k = 1.2$ hr^{-1}. Three days of artificial influent flow and concentration data were generated from sine waves superposed with an Ornstein-Uhlenbeck process (the OU-process can be considered as the continuous version of the discrete first-order auto-regressive process, AR(1)) (Uhlenbeck and Ornstein, 1930). No measurement error was assumed. The time series had an average of 0.1 m^3/s and 16.7 g/m^3, respectively.

For each experiment, three populations were bred, and only the total volume of all reactors was made available to the algorithm. The parameter values used for the evolution are given in Table 5.3. The reactor models of the fittest individual of every population are shown in Figure 5.3. In all breeds of Experiment 1, evolution was stopped because the fitness criterion was met. Notably, the best individuals share the fitness value although their schemes differ. In Experiment 2, in all but one case breeding was stop because the fitness criterion was satisfied. It is now the engineer's task to select the preferred system, taking into account the performance of the system while also considering its simplicity in addition to background knowledge not provided to the GP algorithm.

Lack of identifiability is probably the main challenge when inferring reactor models; both the arrangement of the reactors and the parameter values might be unidentifiable. Under certain circumstances (e.g., a lack of reactions), different arrangements can lead to exactly the same effluent concentrations. This dynamic is visible in the results of the first experiment, in which the arrangement of the reactors cannot be identified. Possible counter-measures include

Table 5.3: The genetic programming parameters used for the experiments (RMSE = root mean square error).

Parameter	Value	Parameter	Value
Population size	500	Tournament size	7
Max. generations	100	Selection probability	0.8
Fitness criterion	$F < 0.001$ g/m^3	Crossover rate	0.50
Parsimony pressure	$\alpha = 10^{-4}$	Mutation rate	0.49
Min. tree depth	5	Reproduction rate	0.01
Max. tree depth	18		

constraining the search space by providing additional knowledge about the system, or using the concentrations within the reactor to construct a more elaborate fitness evaluation. Due to the reactions in the second CSTR of Experiment 2, the arrangement of the reactors was correctly identified.

The fact that in Breed #3 of Experiment 2 breeding was stopped by reaching the maximum number of generations indicates that the optimal solution has not yet been found. It is therefore advisable to investigate the population statistics. Figure 5.4 shows that, although the minimal fitness decreases over time, a significant number of unfit individuals are still being introduced and a high diversity is maintained. A high diversity in the population increases the chance that mutations and crossovers lead to new, superior individuals. Consequently, it would be a matter of time until the global optimum was found if breeding was continued.

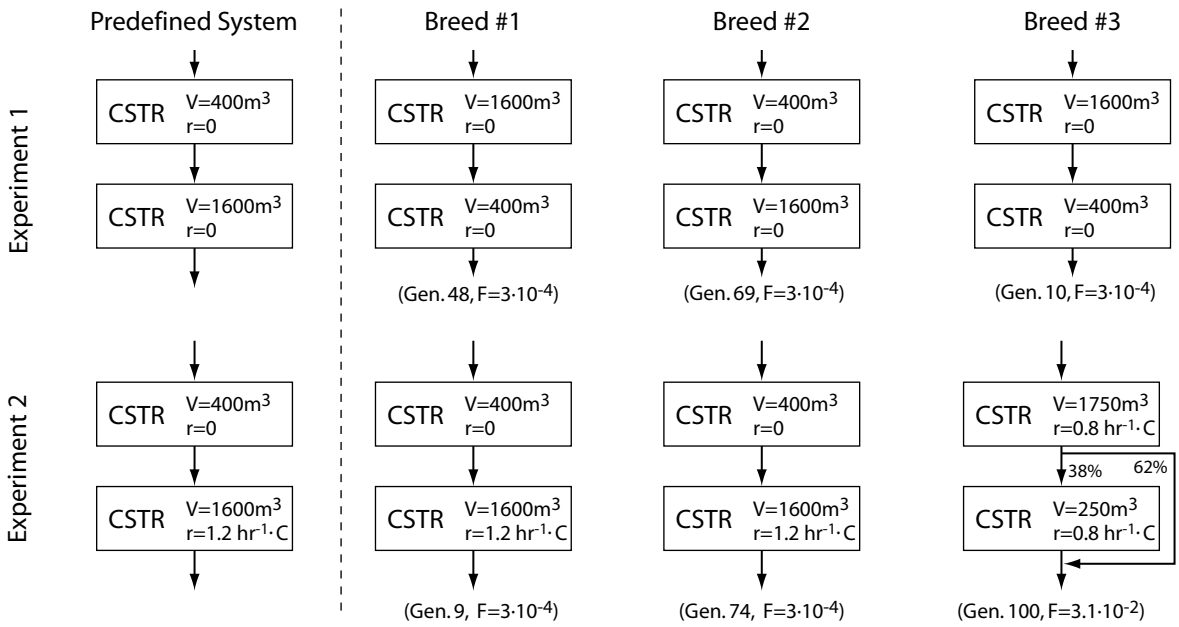

Figure 5.3: Reverse-engineering of two predefined systems (one without and one with reactions). The best individuals of three different breeds are shown, their generation number and RMSE are indicated.

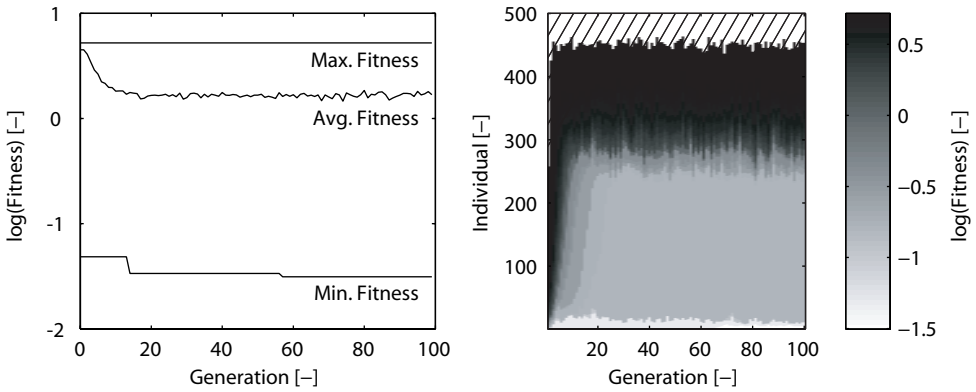

Figure 5.4: Population statistics for Breed #3 of Experiment 2 (Figure 5.3). The minimal, average and maximal fitness is plotted for each generation (left). The evolution of the fitness distribution in the generations is shown (right), the hatched area contains invalid program trees.

5.3.2 Simulation of a tracer experiment

The proposed technique was used to generate a reactor model for a primary clarifier ($V = 750\ \mathrm{m}^3$) in a mid-sized Swiss WWTP and to investigate to which extent virtual tracer experiments performed on the generated models agree with a tracer experiment performed on-site. Influent and effluent temperature was measured for four days, the total WWTP influent flow data was extracted from the process information system of the plant.

Reactor models were obtained in three breeds without constraining the total reactor volume and using the temperature time series as well as the parameters given in Table 5.3. The best-performing models were used to simulate a tracer experiment, the results of which were compared to a tracer experiment performed on site (Figure 5.5).

The evolved systems model the measured temperature data accurately and generally agree with the tracer experiment. Hence, they show that the primary clarifier can be modeled by a relatively simple reactor scheme. Because the modeled total volume corresponds approximately to the real reactor volume, the non-existence of shortcuts or dead zones can be assumed; this is important information that would not be available if modeling was based on experience and intuition only.

5.3.3 Computational time

Although the presented method can easily be run on a modern personal computer with reasonable computational time, there is much room for improvement: *i)* The structure of the model and parameters are optimized simultaneously. A random mutation to the structure is as likely as a change to a parameter value. Because the ERC sets that contain the parameter values are rather large, there can be a large lag before a particular structure co-occurs with a particular parameter value. This lag is a known weakness of GP; ERCs are seen as "the skeleton in the closet of GP methods" (Evett and Fernandez, 1998). However, several strategies to overcome this weakness exist (e.g., Evett and Fernandez, 1998). *ii)* The solution of the ODE system can be very costly for long time series and systems requiring small step sizes. Although system evaluation prior to solving or early discarding strategies could be prescribed, these methods would directly influence the diversity of the population.

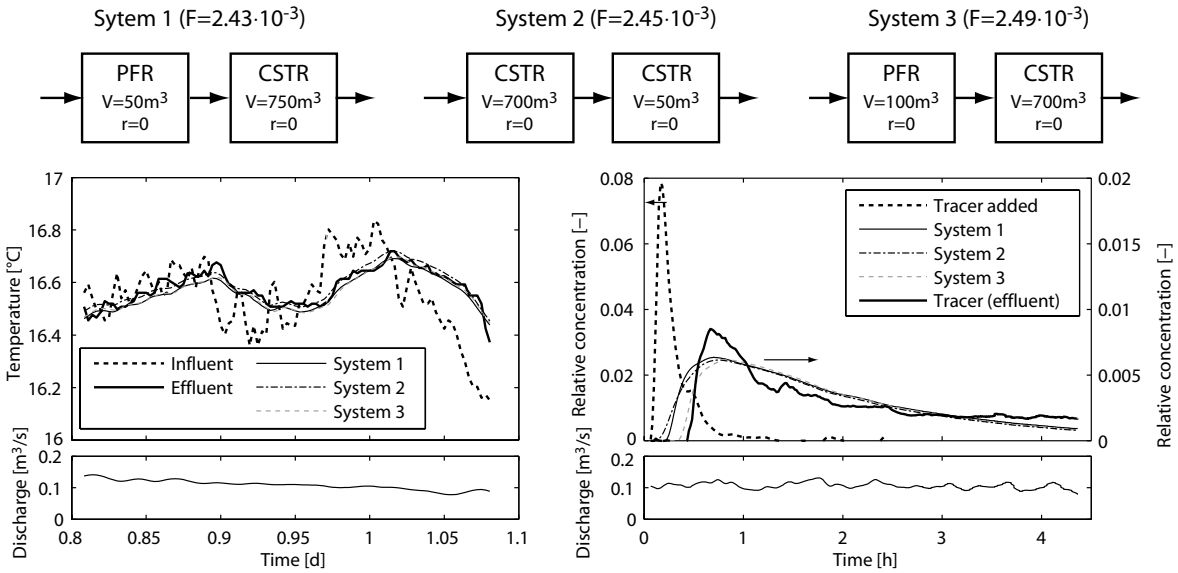

Figure 5.5: The best-performing systems of three populations evolved in parallel for a primary clarifier given influent and effluent temperature data (top). A section of the simulated and measured temperature time series are plotted (bottom left) and a tracer experiment (conservative tracer) performed on-site is compared with tracer responses simulated with the best-performing systems (bottom right).

5.4 Conclusions

In this paper, grammar-based GP was applied to generate hydraulic reactor models. A special tree encoding that can represent hydraulic reactor models as program trees was introduced.

It was shown that, given influent and effluent measurements from a reactor, GP can evolve a reactor model that reproduces the measured effluent series without additional information on the structure of the solution. We suggested evolving several reactor models in different GP runs and let the modeler choose the most appropriate model after taking additional information into account. This latter step is important because a lack of identifiability, poor data quality and limited computing resources all affect the search and can lead to non-optimal solutions.

Although alternative approaches such as tracer experiments are superior to the approach presented here, they have limited applicability in practice. The presented approach, however, is a cost-effective alternative for extracting additional information on transport phenomena without the need for difficult and costly measuring campaigns and thus allows for more accurate modeling of hydraulic processes in practice, which is key to the successful modeling of WWTP.

References

Alex, J., Kolisch, G., Krause, K., 2002. Model structure identification for wastewater treatment simulation based on computational fluid dynamics. Water Science and Technology 45 (4-5), 325–334.

Evett, M., Fernandez, T., 1998. Numeric mutation improves the discovery of numeric constants in genetic programming. In: Koza, J., Banzhaf, W., Chellapilla, K., Deb, K., Dorigo, M., Fogel, D., Garzon, M., Goldberg, D., Iba, H., Riolo, R. (Eds.), Genetic Programming 1998. Proceedings of the Third Annual Conference. Morgan Kaufmann Publishers, San Francisco, CA, USA, pp. 66–71.

Flores, J., Graff, M., 2005. System Identification Using Genetic Programming and Gene Expression Programming. In: Yolum, P., Güngör, T., Gürgen, F., Özturan, C. (Eds.), Computer and Information Sciences - ISCIS 2005. Vol. 3733. Springer, Berlin, Heidelberg, pp. 503–511.

Gujer, W., 2008. Systems analysis for water technology. Springer, Berlin.

Hocine, S., Pibouleau, L., Azzaro-Pantel, C., Domenech, S., 2008. Modelling systems defined by RTD curves. Computers & Chemical Engineering 32 (12), 3112–3120.

Keesman, K. J., 2011. System identification: An introduction. Springer, London.

Keesman, K. J., Stigter, H., 2002. On compartmental modelling of mixing phenomena. In: Proceedings of the 15th IFAC World Congress, L. Basanez, J. A. de la Puente (Eds.), Barcelona.

Koza, J. R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, Mass.

Laquerbe, C., Laborde, J. C., Soares, S., Ricciardi, L., Floquet, P., Pibouleau, L., Domenech, S., 2001. Computer aided synthesis of RTD models to simulate the air flow distribution in ventilated rooms. Chemical Engineering Science 56 (20), 5727–5738.

Ljung, L., 1987. System identification: Theory for the user. Prentice-Hall, Englewood Cliffs, New Jersey.

McKay, B., Willis, M., Barton, G., 1997. Steady-state modelling of chemical process systems using genetic programming. Computers & Chemical Engineering 21 (9), 981–996.

Soule, T., Foster, J. A., 1998. Effects of code growth and parsimony pressure on populations in genetic programming. Evolutionary Computation 6, 293–309.

Tchobanoglous, G., 2003. Wastewater engineering: Treatment and reuse, 4th Edition. McGraw-Hill, Boston, Mass.

Tsakonas, A., 2006. A comparison of classification accuracy of four genetic programming-evolved intelligent structures. Information Sciences 176 (6), 691–724.

Uhlenbeck, G. E., Ornstein, L. S., 1930. On the theory of the Brownian motion. Physical Review 36 (5), 823–841.

Chapter 6

General conclusions and outlook

General conclusions and outlook

6.1 General conclusions

Wastewater treatment plant (WWTP) operation generates a considerable amount of measurement data. In modern plants, the data are carefully archived in the supervisory control and data acquisition system (SCADA system) or in a specially acquired process information system. The data are important for plant control and are used for reporting. However, the use seldom goes beyond this scope; hence, the data are not used optimally. Systematic exploitation of the information hidden in the data that is potentially valuable for plant operation, optimization and modeling is not performed. The main reasons for this are the high dimensionality of the data and, in this context, the lack of appropriate tools for data analysis.

In the past few years, many data mining techniques have emerged that are capable of analyzing massive amounts of data. Efficient data-driven modeling techniques have been developed that are especially suitable whenever the speed of data acquisition is faster than the speed of manual analysis and interpretation, which is typically the case with WWTP data. Consequently, adaptation of these techniques to and application with WWTP data are promising for further exploitation of available data in many potential applications. If practicability and cost-effectiveness are thoroughly considered, the plant operators and engineers alike can be provided with suitable tools for improving plant operation, optimization and modeling.

The applications carried out in collaboration with WWTP staff showed that data mining and data-driven modeling techniques are indeed adequate for the support of WWTP operation if the developed tools can easily be setup and deployed and if their application is comprehensible.

6.1.1 WWTP influent sewage characterization

Although relevant, WWTP operators are generally provided neither detailed information about the sewage producers in the catchment nor the composition of their sewage. In Chapter 3, a two-staged clustering approach was presented that can identify typical sewage compositions by clustering UV/Vis absorption spectra measured at the WWTP inlet. Because a spectrum is unique for a certain sewage composition, it can be considered a fingerprint of that composition. The two-staged clustering approach consists of a self-organizing map (a type of artificial neural network with unsupervised learning that is noise tolerant and copes well with high-dimensional data) to generate a smaller but still representative set of UV/Vis spectra in the first stage. In the second stage, Ward's hierarchical agglomerative clustering algorithm yields a small set of clusters of characteristic compositions.

The clustering model is most useful when particular dischargers can be assigned to clusters. In a full scale experiment, it was possible to assign one of five detected clusters to an industrial

laundry by analyzing the cluster centroids and temporal discharging patterns, and by inducing rules from the comparison of cluster occurrences with other influent measurements. Given the labeled cluster model, newly measured spectra can be classified, and thus, particular discharging events can be detected. For the industrial laundry, 93 out of 95 discharging events measured in a validation experiment were classified correctly. A dilution experiment revealed that successful detection at the plant influent is possible if the dilution contains more than 7% laundry sewage.

Due to the simple setup of the clustering model and the general availability of photospectrometers that cope with the harsh and varying conditions at the WWTP influent, deployment on-site for day-to-day use is feasible. The operator can use the clustering models not only to detect specific discharging events. By monitoring how well the measured spectra fit in the cluster model, unknown and possibly harmful wastewater compositions can also be detected. If the model was connected to the process control system, potentially harmful sewage could immediately be bypassed to a storage basin. In addition, the development of the catchment area can be visualized by routine calibration of a new clustering model and by comparing its number of clusters and cluster centroids with previous models.

6.1.2 Sensor diagnosis and sensor substitution

In Chapter 2 it is shown that with automated data-driven modeling, it is possible to provide the WWTP operator with cost-effective redundant virtual signals of real hardware sensors that are based on measurement data available in the SCADA system. These software sensors can be substituted for maintenance-intensive and failure-prone hardware sensors, or they can be considered for sensor diagnosis.

In two full scale experiments, software sensors were derived with four different data-driven modeling techniques with and without dimensionality reduction and given varying levels of expert knowledge. The modeling techniques were generalized least squares regression (GLSR), artificial neural networks (ANNs), self-organizing maps (SOMs), and random forests (RFs).

The generalization error of the resulting software sensors depended on the chosen modeling technique, if dimensionality reduction was applied and on the level of expert knowledge considered. The non-linear modeling techniques (ANN, SOM and RF) generally performed better than GLSR; dimensionality reduction worsened the performance. Increasing expert knowledge correlated with increasing generalization error due to the lack of variables that might show some local correlations and therefore lead to better accuracy. For long-term predictions, however, models based on higher degrees of expert knowledge clearly performed better.

The data-driven software sensors require a careful model check prior to deployment. This check requires expert knowledge and model transparency, which is only given for GLSR models and SOMs without dimensionality reduction. Both methods provide convenient means for inspecting the model internals. In Chapter 2, prediction intervals or quantification errors that can be considered as indicators for decreasing accuracy, how rare events can be handled and how failures of software-sensor input signals can be detected are discussed.

6.1.3 Operational issues caused by asymmetric discharge distributions

In practice, multiple reactor lanes are often operated in parallel. Hydraulic distribution devices are used to split the flow and provide equalized charging of the individual lanes. These devices, however, are often inaccurate, which leads to uneven loading that can result in performance losses. Quantitative assessment with common methods such as tracer experiments and mobile discharge measurements is not trivial.

In Chapter 4, a practicable new method was presented that estimates the discharge distribution as a function of the total discharge at the flow divider. Based on this function, the operator can, iteratively, perform modifications that affect the discharge distribution and reassess the situation until the distribution is satisfyingly uniform. On the other hand, the distribution function can be used to better estimate the flows to the individual reactor lanes than the assumption of an equal distribution.

The new method is based on dynamic time warping, a distance measure often considered for data mining applications, and has very sparse data requirements. The measurement of a signal that exhibits some variability at the flow divider and at the effluent of each reactor downstream of the divider is sufficient. Temperature sensors have proven to be a good choice; they are commonly available and almost maintenance free. In addition, all parameters for the method can be linked to properties that are generally known or can easily be estimated.

Theoretical analysis revealed that the method has better accuracy for systems with low dispersion, if there are no significant reactions that have an effect on the measured signal and if there are no high-frequency disturbances in the influent signal. The application of the method to synthetic systems confirmed these findings and showed that the coefficient of variation of the root mean square deviation, CV(RMSD), of the flow corrected with the estimated function is between 0.06 and 0.006, whereas the CV(RMSD) of the uncorrected flow (i.e., assuming equal distribution), was 0.11.

In a validation experiment at a grit chamber operated in two parallel lanes, the distribution was successfully expressed as a function of total flow. It was further shown that the choice of the method parameter values is not critical. The measured time series, however, should be sufficiently long (at least three days for the validation experiment).

6.1.4 Modeling the hydraulic processes

Generally, an appropriate description of the hydraulic processes is a key element for successful WWTP modeling. In practice, engineers must often rely on their experience or intuition because conventional methods used to gain insights into reactor hydraulics are labor-intensive and costly.

The method to quantify the discharge distribution that was discussed above and presented in Chapter 4 is clearly valuable for the engineer who wants to model a treatment step that is implemented in parallel lanes. He or she can apply the proposed method to obtain an estimate for the reactor inflow that is more accurate than the assumption of equal distribution.

The implementation presented in Chapter 5 goes one step further. Grammar-based genetic programming (GP) is used to search for hydraulic reactor models. The models are encoded as a program tree with functions to generate continuous stirred-tank reactors, plug flow

reactors and reaction kinetics, and with functions to link the generated reactors. This allows the synthesis of almost arbitrary reactor models. The limitations to these generic building blocks guarantee compatibility with common software for WWTP modeling. To guide the search process, data of some easy-to-measure signals at the influent and the effluent of the reactor to model is sufficient.

Although GP is a global optimization algorithm, multiple runs may yield different equally performing reactor models due to noise and measuring errors in the data, and to theoretical identifiability issues and some stop criteria that keep the computation time reasonable. Therefore, a palette of equally performing models was generated in several runs, of which the modeler can choose the most suitable one. This process is superior to modeling exclusively based on experience or intuition.

The given algorithm was applied to reverse engineer two synthetic systems. The systems were correctly identified except for the reactor arrangement of the system without reactions, which is mathematically unidentifiable. In a second experiment, measured temperature data were used to find a suitable reactor model for a primary clarifier with an unknown exact volume. The resulting model was then used to perform a virtual tracer experiment, which is compared with a tracer experiment performed on-site. GP successfully derived three equally performing models that all had a total volume that corresponded with the expected reactor volume (i.e., shortcuts or dead zone can be excluded). The virtual and real tracer experiments generally agreed; however, the systems found by GP did not model the initial delay accurately.

6.1.5 Model complexity vs. model accuracy

Appropriate model complexity is crucial for safe deployment on-site. Because of the data-driven nature of the considered methods, there is a risk that the models do not describe the important processes appropriately; thus a careful model check is essential. The interpretation of a rather complex model or a model that is opaque, however, is difficult. Therefore, it is important that the designed data mining and data-driven modeling techniques generate models that are transparent and that can be interpreted with limited expert knowledge available, despite possible performance trade-offs.

In Chapter 2, it was shown that for software-sensor generation, basic linear regression models are transparent and thus highly interpretable. However, more powerful non-linear modeling techniques exist that are still interpretable. With the SOMs, such a technique was presented. When applied for regression in the context of software-sensor generation, not only were they more accurate than linear equivalents, they also prevented overfitting by coupling the model structure to calibration data. Moreover, SOMs provided a quality measure for model predictions and allowed convenient visualization of the relations between the regressor variables. SOMs were also considered in Chapter 3 for clustering UV/Vis absorption spectra because of their ability to deal with noisy and high-dimensional data.

Another method to control model complexity was chosen in Chapter 5 to synthesize reactor models. The models generated by genetic programming consisted of a set of interconnected building blocks that most environmental and chemical engineers are familiar with (e.g., continuous stirred-tank reactors and plug-flow reactors). With this foundation, basic interpretability of the models is a given.

6.1.6 Role of expert knowledge

Sufficient expert knowledge is decisive for most applications. In Chapter 2, for instance, the pre-selection of suitable software-sensor input signals is critical for long-term model accuracy, and the model check ascertains that the relevant processes are accounted for by the model; both require expert knowledge. In Chapter 3 the expert must label the clusters of characteristic sewage to make full use of the clustering model, and in Chapter 5, the modeler eventually selects the most promising hydraulic reactor model.

The fact that domain knowledge is essential for method development is accounted for in knowledge discovery process models by the explicit definition of the "Understanding of the Problem" phase (cf. Figure 1.1, page 5) and by taking domain knowledge into account for later "Evaluation of the Discovered Knowledge". However, while method development is knowledge intensive, reliable application on-site is still feasible with limited knowledge if the model complexity is adapted and some means for continuous model checks are given (see, for instance, Chapter 2).

6.1.7 A systematic framework for the exploitation of plant databases

To supply a strategic, target-oriented procedure for a data mining and data-driven modeling project, a knowledge discovery process model was introduced in Section 1.1.2 (page 5). The model was based on the model defined by Cios *et al.* (2007), itself a hybrid of the models proposed by Chapman *et al.* (2000) and Fayyad *et al.* (1996). It consists of six connected highly interactive and iterative phases.

In this section, the process model is extended into a systematic framework for the exploitation of WWTP databases with data mining and data-driven modeling techniques. The major tasks to be performed within each of the six phases and feedback loops are described, and important decisions are discussed. The lessons learned and the experience obtained from the projects performed in the context of this thesis are considered to thoroughly construct this framework. Consequently, relevant text passages that apply to the circumstances discussed are cited where appropriate. In response to the application issues raised in the introduction (Section 1.1.3, page 6), the discussion in this section will show where and to what extent the issues occurred, and it will explain how they could be circumvented or mitigated.

A graphical representation of the framework is given in Figure 6.1; the framework can be viewed as a guideline for knowledge discovery with data mining and data-driven modeling projects. Knowledge can be represented in many forms (e.g., as a model and as a set of rules).

Phase 1: Understanding of the problem domain

It is crucial to start any data mining project with this phase, which specifically includes the exact definition of the project's goals. At first glance, this consideration seems to be trivial. Neglecting this phase, however, may result in the expenditure of significant effort to find the right answers to the wrong question.

The phase can be divided into four main tasks. First, the project goals and success criteria are defined from the perspective of the problem domain, ideally in collaboration with future

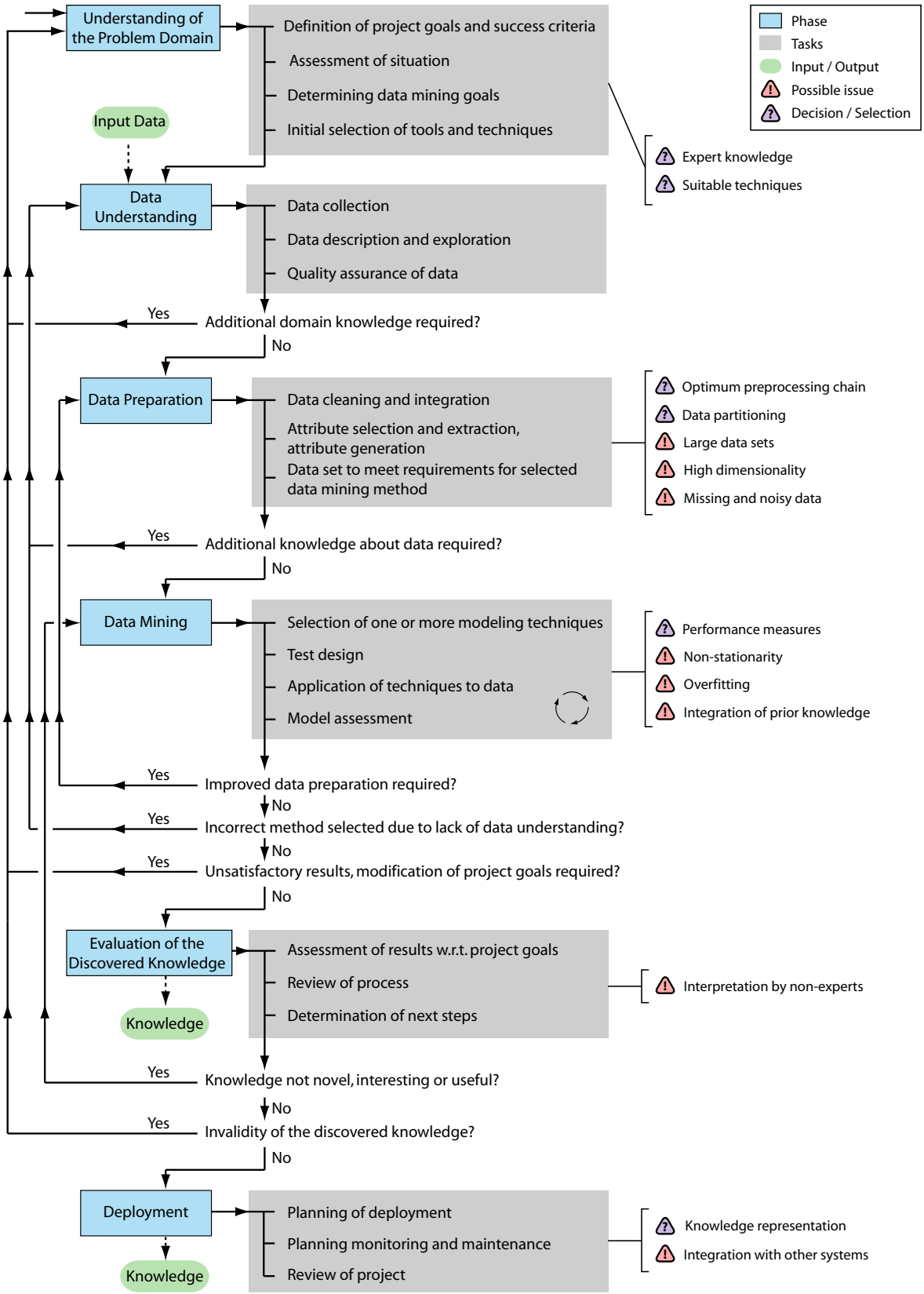

Figure 6.1: Systematic framework for the exploitation of plant databases with data mining and data-driven modeling techniques.

users of the final outcome of the process (i.e., the extracted knowledge). The data miner must balance competing goals, identify constraints and determine factors possibly influencing the outcome. The availability of in-depth expert knowledge is beneficial, specifically for the identification of the WWTP processes involved and the analysis of their dynamics and time scales. Ultimately, available expert knowledge also has implications for the optimum complexity of the outcome (cf. Sections 6.1.5 and 6.1.6, and for detailed information Chapter 2).

The situation is then assessed, i.e., the available resources (including data) are itemized and the assumptions and constraints are investigated in more detail. The consideration of assumptions and constraints shall include a number of aspects of the situation, such as the appropriate complexity of the deployed knowledge.

The third task in this phase consists of the determination of the data mining goal. Specifically, the objectives formulated earlier in the terminology of the problem domain are now translated into the data mining language.

Finally, an initial selection of data mining tools and techniques is made. An early assessment is important because the selection may influence the entire project. However, the selection requires an overview of current data mining techniques, including their field of application, their advantages, and very importantly, their assumptions and limitations. The data mining methods that were applied within this thesis are listed in Table 6.1. The classification introduced in Section 1.1.1 (page 4) is used.

Phase 2: Data understanding

The first task of this phase involves the collection of the data specified in the third task of the previous phase. In the second task of this phase, the collected data are then described and explored to obtain an overview. Data description includes the examination of the basic properties of the data set, such as shape, attribute types and basic attribute statistics. Exploration with the aid of querying and visualization techniques reveals additional information about the attribute distributions and relationships.

The final task of this phase concerns the assurance of data quality. In this task, the applicability of the data to the project objectives is determined. The analysis underlying this determination should consider information on measurement errors, long-term trends, natural variation and altered conditions.

If the data miner realizes that additional domain knowledge is required to (better) understand the data, the feedback loop returning to the “understanding of the problem domain” phase is triggered.

Phase 3: Data preparation

The principal goal of this phase is to make the data available in a form that is compatible with the initially selected data mining tools and techniques. In a data mining project, the probability that this phase is performed more than once is high because the optimum characteristics of the preprocessing chain (i.e., the optimal combination of one or several preprocessing techniques) are not always clear at the outset. Indeed, surveys indicate that

Table 6.1: Main data mining and data-driven modeling techniques discussed in this thesis. These techniques are classified according to their use; particular methods can generally be considered for different problem types. In the rightmost column, section numbers are provided.

Problem type	Generic Method	Specific Method	Reference
Prediction	Regression	Generalized least squares regression (GLSR)	2.2.3 (p 17)
Linear technique, copes with varying variance and autocorrelated residuals			
Prediction	Regression	Artificial neural network (ANN)	2.2.3 (p 18)
Powerful non-linear technique; optimal set-up, however, is highly problem-dependent and model interpretation hindered			
Prediction	Regression	Self-organizing map (SOM)	2.2.3 (p 18)
Non-linear technique, suitable for high-dimensional noisy data, provides means for visualization of non-linear dependencies			
Prediction	Regression	Random forest (RF)	2.2.3 (p 18)
Non-linear regression technique, low risk of overfitting			
Prediction	Classification	Labeled cluster model	3.3.2 (p 39)
Labeled cluster model can be applied to classify new data; labeled Ward clusters in this case			
Description	Clustering	Self-organizing map (SOM)	3.2.3 (p 37)
SOMs can be used to cluster groups of high-dimensional data on a lower dimensional discrete lattice, noise tolerant			
Description	Clustering	Ward clustering	3.2.3 (p 38)
Hierarchical agglomerative clustering algorithm, suitable for non-spherical clusters, does not scale well[a]			
Description	Dependency modeling	Genetic programming (GP)	5.2.2 (p 71)
GP considered for system identification			
Description	Dependency modeling	Dynamic time warping (DTW)	4.2.2 (p 56)
DTW computes a path that describes the warping required between two time series for the best match			
Description	Summarization	Relief	4.4.1 (p 59)
Relief ranks attributes by information content			

[a]In Section 3.2.3 (page 36) SOM and Ward clustering are combined to produce a dual clustering approach that combines the strengths of both methods.

approximately 50% of the time required by a data mining project is invested in this phase (Cios *et al.*, 2007).

The "data preparation" phase should begin with the cleaning and integration of the data. Several algorithms for outlier detection and reconciliation exist for the purpose of data cleaning. Reconciliation is required if the data originate from different sources and must be combined. Subsequently, irrelevant attributes are dropped, existing attributes are transformed or new attributes are generated. Obviously, this process is highly dependent on the specific problem and the selected data mining technique. Hence, it is not possible to give general advice. In Table 6.2, an overview on the preprocessing methods considered in this thesis is given, and the reasons for the utilization of particular methods are briefly explained.

If the selection of the suitable preprocessing methods requires additional or more specific information on the data, the feedback loop to the "data understanding" loop is triggered.

Four possible application issues may require attention during this phase. Preselected data mining algorithms may not scale very well for *large databases or data sets*. Mitigation strategies include sampling, approximation and parallel processing. In addition, more efficient algorithms might be applied. Large data sets were an issue for software-sensor generation in Chapter 2 (high number of attributes), for the assessment of the discharge distribution in Chapter 4 (number of records) and for the synthesis of the hydraulic reactor models in

Table 6.2: Preprocessing techniques presented in this thesis. In the rightmost column, section numbers are provided.

Issue	Cure	Reference
Large data set	Down-sampling	2.2.2 (p 16)
Large data set	Self-organizing map (SOM)	3.2.3 (p 37)
Noisy data	Low-pass filter	4.2.2 (p 56)
Noisy data	Self-organizing map (SOM)	3.2.3 (p 37)
Outliers	TRM filter	2.2.2 (p 16)
High dimensionality	Principal components analysis (PCA)	2.2.2 (p 16)
Dynamics	Approximation by lagging	2.2.2 (p 16)
Non-stationarity	Differencing	2.2.2 (p 16)
Different scaling	Normalization	4.2.2 (p 56)
Misc.	Other transformations (log, empirical formulae, etc.)	2.2.2 (p 16), 3.2.3 (p 36)

Chapter 5 (number of records). The mitigation strategies included efficient variable selection techniques for the first case and parallel processing for the last two. In Chapter 3, self-organizing maps were applied to generate a smaller, yet still representative data set.

High-dimensional data sets have an increased search space. More importantly, however, they may generate meaningless results due to the "curse of dimensionality" (the higher the dimensionality, the more equidistant the data points, cf. Verleysen and François, 2005). This issue emerged in Chapter 3 in relation to the clustering of UV/Vis spectra. However, due to the ability of the selected data mining method (self-organizing maps) to process high-dimensional and noisy data, it was possible to circumvent the "curse of dimensionality". In Chapter 2, principal components analysis (PCA) was applied for dimensionality reduction but was later discarded in favor of stepwise selection due to the limited physical interpretability of the resulting principal components.

Missing and noisy data can be an issue if the data quality is inferior. In this thesis, however, no problems resulted from this potential issue due to good data quality as a result of the use of robust and maintenance-free sensors (Chapters 3, 4 and 5) or due to treatment with low-pass filters, downsampling or the use of robust trimmed repeated median filters (Chapter 2).

Phase 4: Data mining

As the first step in the "data mining" phase, the actual data mining technique is selected. Although the initial selection of tools and techniques made at the beginning of the project was relatively general (for example, hierarchical agglomerative clustering), the selection here is specific (for example, Ward clustering; cf. Table 6.1).

Before the data mining method is applied, it is first desirable to design tests to assess the quality and validity of the outcome. The tests are based on the project goals, formulated in data mining terminology (Phase 1). For supervised techniques, it is possible to define error measures. For regression problems, the coefficient of variation of the root mean squares error was generally considered (cf., Section 2.2.2, page 17). For classification problems, confusion matrices were considered to indicate false positive and false negative rates (Section 3.3.3, page 40). However, for unsupervised techniques such as clustering, alternative indices

are required. For example, an internal criterion can be defined for clustering models that compares the average within-cluster variance to the average variance between clusters (e.g., the Davies-Boudlin index, cf. Section 3.2.3, page 38). To cite an additional example, the self-organizing map allows the calculation of the quantification error (Section 3.2.3, page 37), which says how well a data point is represented in the cluster model, and the calculation of the topographic error, which provides information on the quality of the map (cf. Section 3.3.1, page 38).

For an independent error estimate, it is important not to train and test the model on the same dataset. The dataset should be split it into a calibration set and a validation set. Several data partitioning techniques exist. The choice of techniques depends on the error type of interest to the data miner (cf. Section 2.2.2 on page 16 for an introduction to split-sample validation and cross-validation).

Subsequently, the data mining technique is applied to the data. The model is then assessed with the designed tests. If more than one technique was selected, this tasks would be repeated separately for each technique.

Several situations can trigger a feedback loop:

- Back to phase 1 if the results obtained with the selected method are not satisfactory or if the project goals need to be revised
- Back to phase 2 if a lack of understanding of the data caused the selection of an inappropriate data mining method
- Back to phase 3 if the data preparation needs improvement, often caused by the specific requirements of a data mining method

Overfitting is a major application issue. It is probable this issue will arise in practice. In this situation, the model lacks generalization ability and performs poorly if the model is applied to new data. If the generalization ability of a model is not assessed with an independent test set, the modeler might not realize that overfitting occurred. In this thesis, overfitting could successfully be mitigated by controlling the model complexity (cf. Section 6.1.5), by choosing optimal model structure in cross-validation procedures (cf. Section 2.2.2, page 16) and by considering modeling techniques that intrinsically prevent overfitting (e.g., random forests).

Another issue is *non-stationarity*, a common feature of WWTP data caused by changing environmental conditions and/or process changes. If the data mining task is the detection of these changes (e.g., UV/Vis clustering to visualize the development of the catchment) or if the required data length for analysis is shorter than the typical time scales of the changes (Chapters 4 and 5), non-stationarity is not an issue. For the applications in Chapters 2 and 3, model recalibration was suggested if any significant changes were detected.

The *incorporation of prior knowledge* in the form of domain knowledge for method development is not possible in a simple way for many current methods and tools. In fact, the majority of data mining methods applied in this thesis, did not allow any extensive systematic inclusion of prior knowledge. As an alternative, domain knowledge can often be considered in a relatively straightforward way in the “data preparation” phase, as it was done in Chapter 2 with available expert knowledge.

Phase 5: Evaluation of the discovered knowledge

In contrast to the previous phase where the resulting data mining model was evaluated in view of the data mining goals, the model is now evaluated in the light of the project goals formulated in domain terminology in Phase 1. Any novel and interesting patterns discovered during this phase are noted.

If evaluation is successful, a thorough review of the data mining process is appropriate. This review ensures that no important factor or task has been omitted. The next steps are then chosen, particularly if the findings, in whatever form, are to be deployed.

If the discovered knowledge is invalid due to, e.g., lack of domain understanding and lack of understanding of the problem, the entire project must be repeated. If the discovered knowledge is neither novel nor interesting, the feedback loop that returns the data miner to the "Data Mining" phase is triggered.

Making discovered patterns understandable is often nontrivial and can be a legitimate issue. Making discovered patterns understandable is a principal goal underlying the entire thesis and is strongly connected with the demand for practical applicability of the methods. In addition to the visualization of the discovered knowledge, the safe and reliable use of the knowledge in decision processes requires careful selection of the visualization technique and the inclusion of quality and uncertainty measures to avoid misinterpretation. However, the choice of the most suitable means for communication of the discovered knowledge depends strongly on the specific problem and the method considered.

Phase 6: Deployment

The last phase involves the deployment of the discovered knowledge. First, a strategy for deploying the outcome within the problem domain should be elaborated, followed by the planning of monitoring and maintenance. Monitoring and maintenance are important if the deployed outcome will be part of day-to-day operations and is decisive for the success of the project if the end users do not have in-depth expert knowledge. Eventually, the overall project is reviewed and documented.

The discovered knowledge should be *integrated with other systems* for several reasons. If knowledge has the form of a predictive model, for instance, and is intended to support the day-to-day operation of a WWTP, the model should run independently in the background and retrieve the required data automatically. Otherwise, the results should be fed back to the SCADA system and visualized within the system's user interface. Integration with other systems is not addressed in this thesis (see below).

6.2 Outlook

The rapid progress in information technology and the advancements in data analysis will continue to unleash the potential for even more advanced and thorough utilization of available data. For the field of WWTP operation, however, further developments are required to take full advantage of this potential.

6.2.1 Data management

On-site data management is the process that ensures storage and accessibility of data, which is the principal requirement for data mining and data-driven modeling. Unfortunately, even though the costs for storage have dropped significantly in the last several years, many plants have not kept up and therefore still store their data either in a compacted form, e.g., as daily averages, or do not use long-term archiving. The point is to show the operators that long-term archiving is feasible and that it will, in return, be valuable for plant operation.

Plant operators are experienced addressing various types of incidents that occur during plant operation. From the data miner's perspective, the availability of this information would be helpful for data understanding. At best, these incidents would be recorded in the same database as the process data, e.g., as metadata describing the process data. These metadata significantly augment the information content of the data. In this regard, the development of a simple and well-defined way to record these metadata is envisioned. Here too, the point is to illustrate that the extra effort for gathering the metadata will, in the long run, pay off.

Standardization of data management would further facilitate the provisioning of data mining tools. Unified and open interfaces as well as harmonized database schemes would simplify tool deployment and would encourage benchmarking at a finer level.

6.2.2 Process optimization

The possible further developments for WWTP process optimization are many. Of particular interest are the reduction of consumed resources (e.g., energy) while maintaining constant effluent quality and increasing the effluent quality while maintaining the amount of resources consumed constant. However, with respect to the development of practicable data mining and data-driven modeling techniques, the following two points are critical. First, the knowledge discovery process should be extended with the explicit assessment of a baseline that can be assumed to evaluate the effect of a particular implementation, as part of the "Understanding of the Problem"-phase (cf. Figure 1.1, page 5). Second, the appropriate way to convey the extracted information must be decided upon, i.e., whether it is directly fed to the process control system or visualized to the operator.

Currently available data mining methods have difficulty addressing the non-stationarity and the high dynamics that vary on different time scales, which is typical for wastewater treatment processes. Here, the development of novel methods that inherently consider these features while still being practicable is due.

6.2.3 Deployment strategies

In this thesis, the discussion of deployment strategies focused on technical aspects, such as robustness and long-term accuracy. Obviously, there are many practical aspects as well that are decisive for successful deployment. From the operator's point of view, a seamless integration in his or her SCADA system is desired. The user interfaces must be simple and intuitive yet still provide the desired information at a glance. From the tool provider's perspective, optimal commissioning and maintenance are key factors. Choosing SaaS (software as a service; i.e., the software is hosted on a centralized server platform and is accessible via the internet)

as a delivery method may have decisive advantages (e.g., computational resources can be provided centrally and software maintenance and support is simplified) if the developer is able to circumvent the issues associated with data transfer and security.

6.2.4 Specific suggestions for further research

The following suggestions are specific to the individual chapters of this thesis.

Chapter 2 Data-driven modeling approaches to support WWTP operation

- To account for the different measuring locations of the software-sensor inputs, time-lagged variables were introduced. This is a very rough approximation of the hydraulic processes. The application of the method presented in Chapter 5 to infer hydraulic reactor models could be a promising extension for software-sensor generation to more appropriately model plant hydraulics.
- The use of available prior knowledge is limited to the selection of possible input signals and the restriction of the lags to be considered. A simple way to represent available knowledge in a form compatible with the modeling methods would help the generation of software-sensor models that are based on true rather than on arbitrary correlations.

Chapter 3 Identification of industrial wastewater by clustering UV/Vis spectra

- The implementation of a method to induce rules for the occurrence of a specific wastewater composition based on other signals would allow the application of the presented method in larger catchments where the installation of an UV/Vis probe at the WWTP influent is no longer sufficient and when it is not feasible to mount a UV/Vis probe at the effluent of all relevant subcatchments. Consequently, the installation of other (possibly easier to measure) signals would suffice.
- Trajectories can be plotted on a self-organizing map to visualize temporal transitions. Given a set of labeled historic transitions, the operator could compare the current trajectory, anticipate its continuation and possibly intervene. Obviously, this concept is not limited to clustered UV/Vis spectra.

Chapter 4 Discharge distribution at hydraulic flow dividers

- An approach based on dynamic time warping was presented as a tool to quantify the discharge distribution in hydraulic flow dividers. An intermediate result is the hydraulic residence time in the reactors downstream of the divider or their inflows if the volume is known. Here, it could be tested if the method is viable for estimating the discharge in reactors or pressure mains and the flow velocity in gravity sewers.
- In brief, dynamic time warping estimates the hydraulic residence time by determining the delays of characteristic patterns between the reactor influent and the effluent. Theoretical analysis has shown that dispersion in the reactor and reactions influence the time shift. If the exact value of the discharge is given, the determined delays might provide further insight into the mixing processes and reactions in the reactor.

Chapter 5 Automatic reactor model synthesis with genetic programming

- The current implementation considers one state variable per reactor and one reaction per reactor model. This has proven sufficient for the considered applications that used (synthetic) temperature data. However, the introduction of additional state variables might be required if measuring data of inter-reacting compounds are considered.
- Currently, there are limited possibilities to take into account available prior knowledge and measuring data. Extensive inclusion of these, however, would not only reduce identifiability issues but also decrease the computation time. Therefore, ways to, e.g., specify partial layouts and given fluxes and consider measurements other than those from the influent or effluent must be investigated.

References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0: Step-by-Step Data Mining Guide.

Cios, K. J., Swiniarski, R. W., Pedrycz, W., Kurgan, L. A., Cios, K., Swiniarski, R., Kurgan, L., 2007. The Knowledge Discovery Process. In: Data Mining. Springer, New York, NY, pp. 9–24.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Magazine 17, 37–54.

Verleysen, M., François, D., 2005. The curse of dimensionality in data mining and time series prediction. In: Computational Intelligence and Bioinspired Systems, Lecture Notes in Computer Science 3512. Springer, Berlin, Heidelberg, pp. 758–770.

Appendix A

Software availability

Software availability

The software for Chapters 3, 4 and 5 is available for download at the personal homepage of the author (`http://www.eawag.ch/~duerreda`) or by request via e-mail (`david.duerrenmatt@eawag.ch`). Please note that the software is provided as-is.

A.1 Identification of industrial wastewater by clustering UV/Vis spectra

In Chapter 3, a two-staged clustering approach was introduced to cluster UV/Vis spectra. The software was implemented in MATLAB and is available for download. For installation instructions, please refer to the file `INSTALL.TXT` in the zip archive. A ready-to-run example is provided, including a reduced data set of UV/Vis spectra.

Name	in_cluster
Programming language	MATLAB (Version 6.5 or higher required)
Package dependencies	Statistics Toolbox, SOMTOOLBOX[a)], SOMVIS[b)]
Installation	cf. `INSTALL.TXT`
Package description	cf. `README.TXT`
Example computation	Open and run file `example.m`
Download	`chap3_in_cluster.zip`

[a, b)] The open source package SOMVIS can be downloaded from `http://www.ifs.tuwien.ac.at/dm/somvis-matlab/somvis.zip`, it includes a patched version of SOMTOOLBOX.

A.2 Discharge distribution at hydraulic flow dividers

The algorithm presented in Chapter 4 to compute the discharge distribution at hydraulic flow dividers was implemented in MATLAB and is available for download. Example data is included.

Name	discharge_distribution
Programming language	MATLAB (Version 7.8 or higher required)
Package dependencies	Matlab Toolboxes: Curve Fitting Toolbox, Signal Processing Toolbox, Statistics Toolbox
Installation	Not required
Package description	cf. `README.TXT`
Example computation	Open and run file `main.m`
Download	`chap4_discharge_distribution.zip`

The software is designed to take advantage of parallel computing. Although not required, it is recommended to use the parallel processing features of MATLAB for faster computation. Further details are given in the preamble of file `main.m`.

A.3 Automatic reactor model synthesis with genetic programming

The use of genetic programming to synthesize hydraulic reactor models was presented in Chapter 5. The program was implemented in Python and is available for download.

Depending on the reactor system and the length and resolution of the measuring data given, computation can be resource intensive. Python processes occupy only one core of multicore processors. Because the user normally wants to breed several populations, it is thus suggested to run several processes of the program in parallel (e.g., one per core) in order to take full advantage of the available processing power.

In addition, the program is designed to run on Amazon's elastic computation cloud (EC2, `http://aws.amazon.com`) using StarCluster (`http://web.mit.edu/stardev/cluster`). The required plugins are included in the program package.

Name	hydra_gen
Programming language	Python (Version 2.6.x required)
Package dependencies[a)]	numpy, scipy, matplotlib, mpl_toolkit, pysqlite2, psyco, optparse, configparser, pydot, logging, pystep[b)]
Installation	Instructions given in `INSTALL.TXT`
Package description	cf. `README.TXT`
Example computation	Follow instructions given in `README.TXT`
Download	`chap5_hydra_gen.zip`

[a)] The packages are available in the Python package index (`http://pypi.python.org`) and can mostly be installed with the `easy_install` command of the `setuptools` package.

[b)] A patched version of `pystep` with additional visualization capabilities is already contained in this package.

Appendix B

Supporting information for Chapter 2

Supporting information for Chapter 2

This appendix contains supporting information for Chapter 2: "Data-driven modeling approaches to support WWTP operation".

B.1 Modeling techniques

B.1.1 Generalized least squares regression (GLSR)

Generalized least squares regression is a linear modeling technique. In contrast with the ordinary least squares (OLS) estimation method, the GLS estimation does not assume that the variance-covariance matrix of the errors, ε, has the form $\mathrm{Cov}(\varepsilon)_{OLS} = \sigma^2 \boldsymbol{I}$ with the identity matrix $\boldsymbol{I}$, but rather allows $\mathrm{Cov}(\varepsilon)_{GLS} = \sigma^2 \boldsymbol{\Sigma_\varepsilon}$ with a known matrix $\boldsymbol{\Sigma_\varepsilon}$ (Montgomery *et al.*, 2006). This is an important feature that helps cope with the auto-correlated residuals that typically occur when dealing with time series data and incomplete models (Dellana and West, 2009). If the OLS was nevertheless applied, the estimator would still be unbiased, but it is no longer a minimum-variance estimator (Montgomery *et al.*, 2006), which has implications for standard tests and could lead to misleading test scores. Even if the matrix $\boldsymbol{\Sigma_\varepsilon}$ is not known directly, it can be estimated given a suitable restrictive parameterization of $\boldsymbol{\Sigma_\varepsilon}$ (Greene, 2000).

There are good reasons to build the GLS model with only a subset of the available regressor variables. A parsimonious model structure has higher interpretability, and sometimes the prediction accuracy can be improved by removing some variables (Hastie *et al.*, 2009). Because best-subset selection, which basically tries all possible combinations of regressor variables, is not feasible for a high number of variables (Hastie *et al.*, 2009), backward-elimination is applied: first, a model with all N regressor variables is calibrated and its performance is computed. Then, the N possible models with $N-1$ regressor variables are fitted and the best performing one is selected. Now starting from the selected model, all possible models with one regressor variable less are fitted and likewise, the best model is selected. This is repeated until a model results with only one regressor variable. As a result, the backward-elimination procedure yields a path through all possible subsets of which the model with best performance is eventually selected.

The performance criterion applied in this study is the BIC (Bayesian information criterion). The BIC takes into account the quality of the fit, but penalizes complex models (the BIC criterion is similar to the Akaike information criterion (AIC), however, the penalization of model complexity is heavier) (Hastie *et al.*, 2009).

B.1.2 Artificial neural network (ANN)

Artificial neural network is a popular supervised non-linear statistical data modeling tool. The term neural network, however, encompasses a large class of models (Hastie *et al.*, 2009). The feed-forward ANN considered in this paper is a multilayer perceptron (MLP) with one hidden layer; it has an input, a hidden and an output layer. The input layer has a neuron for every input variable. Each neuron is connected with every neuron in the hidden layer. Eventually, each neuron of the hidden layer is connected with the output neuron, whose output is the model response. A neuron is a computational unit that receives one or more inputs and produces an output, depending on the activation function of the neuron. Every input to a neuron is given a weight, and the output of the neuron is the sum of the inputs multiplied by their weights and passed through an activation function. The weights are trained with a back-propagation learning algorithm (Hastie *et al.*, 2009) and early-stopping is applied to prevent over-fitting. In early-stopping, a part of the dataset is not considered for ANN training, but to monitor the performance during training. Training is continued unless the performance of the current iteration is permanently lower than the best-so-far performance over a pre-defined number of iterations. In this study, 15% of the data were retained and the number of iterations was set to 15.

The choice of the optimal structure of the ANN is rather difficult. While one hidden layer is sufficient for function approximation (Cybenko, 1989; Hornik *et al.*, 1989), the number of hidden neurons is problem-dependent and typically in the range of 5-100 (Hastie *et al.*, 2009). Too many hidden neurons might lead to over-fitting, but too few units might result in excessive bias in the outputs (Himmelblau, 2008).

In this study, the optimum number of hidden neurons is evaluated for each experiment individually by training ANNs with varying number of hidden neurons and estimating their generalization errors in a ten-fold cross validation procedure. Eventually, the model structure that minimizes the generalization error is selected.

B.1.3 Self-organizing maps (SOM)

Self-organizing maps are a variant of artificial neural networks based on unsupervised learning, originally proposed by Kohonen (2001). They learn to cluster groups of similar input data in a non-linear projection from a high-dimensional data-space onto a lower-dimensional discrete lattice of neurons on an output layer, called feature map, in an orderly fashion (Céréghino and Park, 2009; Kalteh *et al.*, 2008). This is done in a topology-preserving way, which means that neurons physically located close to each other have similar input patterns. Additionally, SOMs are tolerant to noise, which is especially helpful when dealing with experimental measurements.

Each neuron i has assigned a prototype vector w_i having the same dimensionality as the input data. If a SOM is used for prediction, w_i is the concatenation of the response variable y and the K regressor variables x_k, $w_i = (y_i, x_{i,1}, \ldots, x_{i,K})^T$. During training, these vectors are optimized to represent the complete set of input data; the set of prototype vectors is therefore representative for the data set. The optimization of the prototype vectors is proportional to a learning rate and a neighborhood function, both of which decrease monotonically during the ordering process. The former is a scalar; the latter forms a smoothing kernel around the

prototype vector and verifies that only input vectors within a certain neighborhood affect the prototype vector.

For prediction, the first component of the prototype vector of a trained SOM that has minimum distance to an input vector with an unknown first component is extracted. The distance between the input vector and the prototype vector that is closest (the best-matching unit, BMU) is called the quantification error, *q.e.* It expresses how well an input vector is represented in the SOM and can thus later be considered as a means for software-sensor self-diagnosis (i.e., the higher the *q.e.*, the more uncertain the prediction).

The means to visualize the relations between the variables are the component planes, which are basically cross-sections through the prototype vectors of the feature map. However, if there are variables, it is helpful to quantify their importance. The measure of topological relevance (MTR) is such a measure (Corona *et al.*, 2008).

The models presented in this paper all have a two-dimensional, hexagonal feature map. The number of neurons is determined as $5\sqrt{N}$ (Vesanto and Alhoniemi, 2000), where N is the number of samples, and the ratio of the side lengths was set equal to the two maximum eigenvalues of the data (Park *et al.*, 2006). As a learning rate, a linear function decreasing from 0.05 to 0.01 over 100 iterations is chosen as well as a linearly decreasing neighborhood radius from a radius that covers 2/3 of the distances of the map units to only the winning unit (which is reached after 1/3 of the iterations). Implementation details are given in Wehrens and Buydens (2007).

B.1.4 Random forest (RF)

Random forests is a widely applied machine-learning technique (Mouton *et al.*, 2011; Verikas *et al.*, 2011). RFs are non-linear ensemble classifiers that build on a large collection of B classification or regression trees that are aggregated (Breiman, 2001). To grow a tree, however, a bootstrap sub-sample of size n of the available samples N is taken. In addition, nodes are created by selecting the best regressor variable of k randomly chosen variables from all variables K. In a random forest for regression, the response is the averaged response of all trees.

The RF technique has the advantage that it performs remarkably well with very little tuning required (Hastie *et al.*, 2009) and is not prone to over-fitting (Breiman, 2001; Hastie *et al.*, 2009); hence, it is suitable for highly automated data-driven modeling approaches.

The relative importance of the regressor variables can be measured with samples not selected in the bootstrap sub-samples used to construct a tree. First, the prediction accuracy is recorded by passing these samples down the tree. Then the values of the j-th variable in these samples is permutated and again passed down the tree. The decrease in accuracy, averaged over all of the B trees, is a measure of the importance of variable j (Hastie *et al.*, 2009; Verikas *et al.*, 2011).

In this study, the parameter values are $B = 500$, $k = K/3$ and $n = \log_2(N) + 1$.

References

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Céréghino, R., Park, Y.-S., 2009. Review of the Self-Organizing Map (SOM) approach in water resources: Commentary. Environmental Modelling & Software 24 (8), 945–947.

Corona, F., Reinikainen, S.-P., Aaljoki, K., Perkiö, A., Liitiäinen, E., Baratti, R., Simula, O., Lendasse, A., 2008. Wavelength selection using the measure of topological relevance on the self-organizing map. Journal of Chemometrics 22 (11-12), 610–620.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems 2 (4), 303–314.

Dellana, S. A., West, D., 2009. Predictive modeling for wastewater applications: Linear and nonlinear approaches. Environmental Modelling & Software 24 (1), 96–106.

Greene, W. H., 2000. Econometric analysis, 4th Edition. Prentice Hall Internat., Upper Saddle River, NJ.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction, 2nd Edition. Springer series in statistics. Springer, New York, NY.

Himmelblau, D. M., 2008. Accounts of Experiences in the Application of Artificial Neural Networks in Chemical Engineering. Industrial & Engineering Chemistry Research 47 (16), 5782–5796.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359–366.

Kalteh, A. M., Hiorth, P., Bemdtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling & Software 23 (7), 835–845.

Kohonen, T., 2001. Self-organizing maps, 3rd Edition. Springer, Berlin, London.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2006. Introduction to linear regression analysis, 4th Edition. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, NJ.

Mouton, A. M., Alcaraz-Hernández, J. D., Baets, B. d., Goethals, P. L. M., Martínez-Capel, F., 2011. Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. Environmental Modelling & Software 26 (5), 615–622.

Park, Y.-S., Tison, J., Lek, S., Giraudel, J.-L., Coste, M., Delmas, F., 2006. Application of a self-organizing map to select representative species in multivariate analysis: A case study determining diatom distribution patterns across France. Ecological Informatics 1 (3), 247–257.

Verikas, A., Gelzinis, A., Bacauskiene, M., 2011. Mining data with random forests: A survey and results of new tests. Pattern Recognition 44 (2), 330–349.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11 (3), 586–600.

Wehrens, R., Buydens, L. M. C., 2007. Self- and super-organizing maps in R: The kohonen package. Journal of Statistical Software 21 (5), 1–19.